

Finite Open-World Query Answering with Number Restrictions (Extended Version)

Antoine Amarilli

Télécom ParisTech; Institut Mines–Télécom; CNRS LTCI
antoine.amarilli@telecom-paristech.fr

Michael Benedikt

University of Oxford
michael.benedikt@cs.ox.ac.uk

May 19, 2015

Open-world query answering is the problem of deciding, given a set of facts, conjunction of constraints, and query, whether the facts and constraints imply the query. This amounts to reasoning over all instances that include the facts and satisfy the constraints. We study *finite open-world query answering* (FQA), which assumes that the underlying world is finite and thus only considers the *finite* completions of the instance. The major known decidable cases of FQA derive from the following: the guarded fragment of first-order logic, which can express referential constraints (data in one place points to data in another) but cannot express number restrictions such as functional dependencies; and the guarded fragment with number restrictions but on a signature of arity only two. In this paper, we give the first decidability results for FQA that combine both referential constraints and number restrictions for arbitrary signatures: we show that, for unary inclusion dependencies and functional dependencies, the finiteness assumption of FQA can be lifted up to taking the finite implication closure of the dependencies [8]. Our result relies on new techniques to construct finite universal models of such constraints, for any bound on the maximal query size.

I. Introduction

A longstanding goal in computational logic is to design logical languages that are both decidable and expressive. One approach is to distinguish integrity constraints and queries, and have separate languages for them. We would then seek decidability of the *query answering with constraints* problem: given a query q , a conjunction of constraints Σ , and a finite instance I , determine which answers to q are certain to hold over any instance I' that extends I and satisfies Σ . This problem is often called *open-world query answering*. It is fundamental for deciding query containment under constraints, querying

in the presence of ontologies, or reformulating queries with constraints. Thus it has been the subject of intense study within several communities for decades (e.g. [11, 5, 3, 15, 10]).

In many cases (e.g., in databases) the instances I' of interest are the finite ones, and hence we can define *finite open-world query answering* (denoted here as FQA), which restricts the quantification to *finite* extensions I' of I . In contrast, by *unrestricted open-world query answering* (UQA) we refer to the problem where I' can be either finite or infinite. Generally the class of queries is taken to be the conjunctive queries (CQs) — queries built up from relational atoms via existential quantification and conjunction. We will restrict to CQs here, and thus omit explicit mention of the query language, focusing on the constraint language.

A first constraint class known to have tractable open-world query answering problems are *inclusion dependencies* (IDs) — constraints of the form, e.g., $\forall xyz R(x, y, z) \rightarrow \exists vw S(z, v, w, y)$. The fundamental results of Johnson and Klug [11] and Rosati [18] show that both FQA and UQA are decidable for ID and that, in fact, they coincide. When this happens, the constraints are said to be *finitely controllable*. These results have been generalized by Bárány et al. [3] to a much richer class of constraints, the guarded fragment of first-order logic.

However, those results do not cover a second important kind of constraints, namely *number restrictions*, which express, e.g., uniqueness. We represent them by the class of *functional dependencies* (FDs) — of the form $\forall \mathbf{xy} (R(x_1, \dots, x_n) \wedge R(y_1, \dots, y_n) \wedge \bigwedge_{i \in L} x_i = y_i) \rightarrow x_r = y_r$. The implication problem (does one FD follow from a set of others) is decidable for FDs, and coincides with implication restricted to finite instances [1]. Trivially, the FQA and UQA problems are also decidable for FDs alone, and coincide.

Trying to combine IDs and FDs makes both UQA and FQA undecidable in general [5]. However, UQA is known to be decidable when the FDs and the IDs are *non-conflicting* [11, 5]. Intuitively, this condition guarantees that the FDs can be ignored, as long as they hold on the initial instance I , and one can then solve the query answering problem by considering the IDs alone. But the non-conflicting condition only applies to UQA and not to FQA. In fact it is known that even for very simple classes of IDs and FDs, including non-conflicting classes, FQA and UQA do not coincide. Rosati [18] showed that FQA is undecidable for non-conflicting IDs and FDs (indeed, for IDs and keys, which are less rich than FDs).

Thus a general question is to what extent these classes, FDs and IDs, can be combined while retaining decidable FQA. The only decidable cases impose very severe requirements. For example, the constraint class of “single KDs and FKs” introduced in [18] has decidable FQA, but such constraints cannot model, e.g., FDs which are not keys. Further, in contrast with the general case of FDs and IDs, single KDs and FKs are always finitely controllable, which limits their expressiveness. Indeed, we know of no tools to deal with FQA for non-finitely-controllable constraints on relations of arbitrary arity.

A second decidable case is where all relation symbols and all subformulas of the constraints have arity at most two. In this context, results of Pratt-Hartmann [15] imply the decidability of both FQA and UQA for a very rich non-finitely-controllable sublogic of first-order logic. For some fragments of this arity-two logic, the complexity of FQA has recently been isolated by Ibáñez-García et al. [10]. Yet these results do not apply to arbitrary arity signatures.

The contribution of this paper is to provide the first result about finite query answering for non-finitely-controllable IDs and FDs over relations of arbitrary arity. As the problem is undecidable in general, we must naturally make some restriction. Our choice is to limit to *Unary* IDs (UIDs), which export only one variable: for instance, $\forall xyz R(x, y, z) \rightarrow \exists w S(w, x)$. UIDs and FDs are an interesting class to study because they are not finitely controllable, and allow the modeling, e.g., of single-attribute foreign keys, a common use case in database systems. The decidability of UQA for UIDs and FDs is known because they are always non-conflicting. In this paper, we show that finite query answering is

decidable for UIDs and FDs, and obtain tight bounds on its complexity.

The idea is to *reduce the finite case to the unrestricted case*, but in a more complex way than by finite controllability. We make use of a technique originating in Cosmadakis et al. [8] to study finite implication on UIDs and FDs: the *finite closure* operation which takes a conjunction of UIDs and FDs and determines exactly which additional UIDs and FDs are implied over finite instances. Rosati [17] and Ibáñez-García [10] make use of the closure operation in their study of constraint classes over schemas of arity two. They show that finite query answering for a query q , instance I , and constraints Σ reduces to unrestricted query answering for I , q , and the finite closure Σ' of Σ . In other words, the closure construction which is sound for implication is also sound for query answering.

We show that the same general approach applies to arbitrary arity signatures, with constraints being UIDs and FDs. Our main result thus reduces finite query answering to unrestricted query answering, for UIDs and FDs in arbitrary arity:

Theorem I.1. *For any finite instance I , conjunctive query q , and constraints Σ consisting of UIDs and FDs, the finite open-world query answering problem for I, q under Σ has the same answer as the unrestricted open-world query answering problem for I, q under the finite closure of Σ .*

Using the known results about the complexity of UQA for UIDs, we isolate the precise complexity of finite query answering with respect to UIDs and FDs, showing that it matches that of UQA:

Corollary I.2. *The combined complexity of the finite open-world query answering problem for UIDs and FDs is NP-complete, and it is PTIME in data complexity (that is, when the constraints and query are fixed).*

Our proof of Theorem I.1 is quite involved, since dealing with arbitrary arity models introduces many new difficulties that do not arise in the arity-two case or in the case of IDs in isolation. We borrow and adapt a variety of techniques from prior work: using k -bounded simulations to preserve small acyclic CQs [10], dealing with UIDs following a topological sort [8, 10], performing a chase that reuses sufficiently similar elements [18], and taking the product with groups of large girth to blow up cycles [14]. However, we must also develop some new infrastructure to deal with number restrictions in an arbitrary arity setting: distinguishing between so-called *dangerous* and *non-dangerous* positions when chasing, constructing realizations for relations in a *piecewise* manner following the FDs, reusing elements in a *combinatorial* way that shuffles them to avoid violating the higher-arity FDs, and a new notion of *mixed product* to blow cycles up while preserving fact overlaps to avoid violating the higher-arity FDs.

Paper structure. The general scheme, presented in Section III, is to construct models of UIDs and FDs that are universal up to a certain query size k , which we call *k-universal models*. We start with only *unary* FDs (UFDs) and *acyclic* CQs (ACQs), and by assuming that the UIDs and UFDs are *reversible*, a condition inspired by the finite closure construction.

As a warm-up, Section IV proves the weakened result for a much weaker notion than k -universality, starting with binary signatures and generalizing to arbitrary arity. We extend the result to k -universality in Section V, maintaining a k -bounded simulation to the chase, and performing *thrifty* chase steps that reuse sufficiently similar elements without violating UFDs. We also rely on a structural observation about the chase under UIDs (Theorem V.11). Section VI eliminates the assumption that dependencies are reversible, by partitioning the UIDs into classes that are either reversible or trivial, and satisfying successively each class following a certain ordering.

We then generalize our result to higher-arity (non-unary) FDs in Section VII. This requires us to define a new notion of thrifty chase steps that apply to instances with many ways to reuse elements; the

existence of these instances relies on a combinatorial construction of models of FDs with a high number of facts but a small domain (Theorem VII.7). Last, in Section VIII, we apply a cycle blowup process to the result of the previous constructions, to go from acyclic to arbitrary CQs through a product with acyclic groups. The technique is inspired by Otto [14] but must be adapted to respect FDs.

Complete proofs of our results are provided in the appendix.

II. Background

Instances. We assume an infinite countable set of *elements* (or *values*) a, b, c, \dots and *variable names* x, y, z, \dots . A *schema* σ consists of *relation names* (e.g., R) with an *arity* (e.g., $|R|$) which we assume is ≥ 1 . Following the unnamed perspective, the set of *positions* of R is $\text{Pos}(R) := \{R^i \mid 1 \leq i \leq |R|\}$, and we define $\text{Pos}(\sigma) := \bigsqcup_{R \in \sigma} \text{Pos}(R)$. We identify R^i and i when no confusion can result.

A relational *instance* (or *model*) I of σ is a set of *ground facts* of the form $R(\mathbf{a})$ where R is a relation name and \mathbf{a} an $|R|$ -tuple of values. The *size* $|I|$ of an instance I is its number of facts. The *active domain* $\text{dom}(I)$ of I is the set of the elements which appear in I . For any position $R^i \in \text{Pos}(\sigma)$, we define the *projection* $\pi_{R^i}(I)$ of I to R^i as the set of the elements of $\text{dom}(I)$ that occur at position R^i in I . For $L \subseteq \text{Pos}(R)$, the projection $\pi_L(I)$ is a set of $|L|$ -tuples defined analogously; for convenience, departing from the unnamed perspective, we index those tuples by the positions of L . A *superinstance* of I is a (not necessarily finite) instance I' such that $I \subseteq I'$.

A *homomorphism* from an instance I to an instance I' is a mapping $h : \text{dom}(I) \rightarrow \text{dom}(I')$ such that, for every fact $F = R(\mathbf{a})$ of I , the fact $h(F) := R(h(a_1), \dots, h(a_{|R|}))$ is in I' .

Constraints. We consider integrity constraints (or *dependencies*) which are special sentences of first-order logic. As usual in the relational setting, we do not allow function symbols. The definition of an instance I satisfying a constraint Σ , written $I \models \Sigma$, is standard.

An *inclusion dependency* ID is a sentence of the form $\tau : \forall \mathbf{x} R(x_1, \dots, x_n) \rightarrow \exists \mathbf{y} S(z_1, \dots, z_m)$, where $\mathbf{z} \subseteq \mathbf{x} \cup \mathbf{y}$ and no variable occurs twice in \mathbf{z} . The *exported variables* are the variables of \mathbf{x} that occur in \mathbf{z} , and the *arity* of the dependency is the number of such variables. This work only studies *unary inclusion dependencies* (UIDs) which are the IDs with arity 1. If τ is a UID, we write τ as $R^p \subseteq S^q$, where R^p and S^q are the positions of $R(\mathbf{x})$ and $S(\mathbf{z})$ where the exported variable occurs. For instance, the UID $\forall xy R(x, y) \rightarrow \exists z S(y, z)$ is written $R^2 \subseteq S^1$. We assume without loss of generality that there are no *trivial* UIDs of the form $R^p \subseteq R^p$.

We say that a conjunction Σ_{UID} of UIDs is *transitively closed* if it is closed under implication by the *transitivity rule*: if $R^p \subseteq S^q$ and $S^q \subseteq T^r$ are in Σ_{UID} , then so is $R^p \subseteq T^r$ unless it is trivial. The transitive closure of Σ_{UID} can clearly be computed in PTIME in Σ_{UID} , and it contains all non-trivial UIDs implied by Σ_{UID} over finite or unrestricted instances [7]. We say a UID $\tau : R^p \subseteq S^q$ is *reversible* relative to Σ_{UID} if both τ and its *reverse* $\tau^{-1} := S^q \subseteq R^p$ are in Σ_{UID} .

A *functional dependency* FD is a sentence of the form $\phi : \forall \mathbf{x} \mathbf{y} (R(x_1, \dots, x_n) \wedge R(y_1, \dots, y_n) \wedge \bigwedge_{R^l \in L} x_l = y_l) \rightarrow x_r = y_r$, where $L \subseteq \text{Pos}(R)$ and $R^r \in \text{Pos}(R)$. For brevity, we write ϕ as $R^L \rightarrow R^r$. We call ϕ a *unary functional dependency* UFD if $|L| = 1$; otherwise it is *higher-arity*. For instance, $\forall xx'yy' R(x, x') \wedge R(y, y') \wedge x' = y' \rightarrow x = y$ is a UFD, and we write it $R^2 \rightarrow R^1$. We assume that $|L| > 0$, i.e., we do not allow nonstandard or degenerate FDs. We call ϕ *trivial* if $R^r \in R^L$, in which case ϕ always holds. Two facts $R(\mathbf{a})$ and $R(\mathbf{b})$ *violate* a non-trivial FD ϕ if $\pi_L(\mathbf{a}) = \pi_L(\mathbf{b})$ but $a_r \neq b_r$.

The *key dependency* $\kappa : R^L \rightarrow R$, for $L \subseteq \text{Pos}(R)$, is the conjunction of FDs $R^L \rightarrow R^r$ for all $R^r \in \text{Pos}(R)$; it is *unary* if $|L| = 1$. If κ holds, we call L a *key* (or *unary key*) of R .

Queries. An *atom* $A = R(\mathbf{t})$ consists of a relation name R and a $|R|$ -tuple \mathbf{t} of variables or constants. A *conjunctive query* CQ is an existentially quantified conjunction of atoms. In this paper we focus for simplicity on Boolean queries (queries without free variables), but all our results hold for non-Boolean queries as well, by the standard method of enumerating the assignments. The *size* $|q|$ of a CQ q is its number of atoms.

A *Berge cycle* in a Boolean CQ q is a sequence $A_1, x_1, A_2, x_2, \dots, A_n, x_n$ with $n \geq 2$, where the A_i are pairwise distinct atoms of q , the x_i are pairwise distinct variables of q , and x_i occurs in A_i and A_{i+1} for $1 \leq i \leq n$ (with addition modulo n , so x_n occurs in A_1). We call q *acyclic* if q has no Berge cycle and if no variable of q occurs more than once in the same atom. We write ACQ for the class of acyclic CQs.

A Boolean CQ q *holds* in an instance I exactly when there is a homomorphism h from the atoms of q to I such that h is the identity on the constants of q (we call this a *homomorphism from q to I*). The image of h is called a *match* of q in I .

QA problems. We define the *unrestricted open-world query answering* problem (UQA) as follows: given a finite instance I , a conjunction of constraints Σ , and a Boolean CQ q , decide whether there is a superinstance of I that satisfies Σ and violates q . If there is none, we say that I and Σ *entail* q and write $(I, \Sigma) \models_{\text{unr}} q$.

This work focuses on the *finite query answering problem* (FQA), which is the variant of open-world query answering where we require the counterexample superinstance to be finite; if none exists, we write $(I, \Sigma) \models_{\text{fin}} q$. Of course $(I, \Sigma) \models_{\text{unr}} q$ implies $(I, \Sigma) \models_{\text{fin}} q$. We say a conjunction of constraints Σ is *finitely controllable* if FQA and UQA coincide: for every finite instance I and every Boolean CQ q , $(I, \Sigma) \models_{\text{unr}} q$ iff $(I, \Sigma) \models_{\text{fin}} q$.

The *combined complexity* of the UQA and FQA problems, for a fixed class of constraints, is the complexity of deciding it when all of I , Σ (in the constraint class) and q are given as input. The *data complexity* is defined by assuming that Σ and q are fixed, and only I is given as input.

Chase. We say that a superinstance I' of an instance I is *universal* for constraints Σ if $I' \models \Sigma$ and if for any CQ q , $I' \models q$ iff $(I, \Sigma) \models_{\text{unr}} q$. We now recall the definition of the *chase* [1, 13], a standard construction of (generally infinite) universal superinstances. We assume that we have fixed an infinite set \mathcal{N} of *nulls* which is disjoint from $\text{dom}(I)$. We only define the chase for transitively closed UIDs, which we call the *UID chase*.

We say that a fact $F_a = R(\mathbf{a})$ of an instance I is an *active fact* for a UID $\tau : R^p \subseteq S^q$ if, writing $\tau : \forall \mathbf{x} R(\mathbf{x}) \rightarrow \exists \mathbf{y} S(\mathbf{z})$, there is a homomorphism from $R(\mathbf{x})$ to F_a but no such homomorphism can be extended to a homomorphism from $\{R(\mathbf{x}), S(\mathbf{z})\}$ to I . In this case we say that a_p *wants* to occur at position S^q in I , written $a_p \in \text{Wants}(I, S^q)$, and that we *want* to apply the UID τ to a_p , written $a_p \in \text{Wants}(I, \tau)$. Note that $\text{Wants}(I, \tau) = \pi_{R^p}(I) \setminus \pi_{S^q}(I)$.

The result of a *chase step* on the active fact F_a for τ in I (we call this *applying τ to F_a*) is the superinstance I' of I obtained by adding a new fact $F_n = S(\mathbf{b})$ defined as follows: we set $b_q := a_p$, which we call the *exported element* (and S^q the *exported position* of F_n), and use fresh nulls from \mathcal{N} to instantiate the existentially quantified variables of τ and complete F_n ; we say the corresponding elements are *introduced* at F_n . This ensures that F_a is no longer an active fact in I' for τ .

A *chase round* of a conjunction Σ_{UID} of UIDs on I is the result of applying simultaneous chase steps on all active facts for all UIDs of Σ_{UID} , using distinct fresh elements. The UID chase $\text{Chase}(I, \Sigma_{\text{UID}})$ of I by Σ_{UID} is the (generally infinite) fixpoint of applying chase rounds. It is a universal superinstance for Σ_{UID} [9].

As we are chasing by transitively closed UIDs, if we perform the *core chase* [13] rather than the

UID chase defined above, we can ensure the following *Unique Witness Property*: for any element $a \in \text{dom}(\text{Chase}(I, \Sigma_{\text{UID}}))$ and position R^p of σ , if two different facts of $\text{Chase}(I, \Sigma_{\text{UID}})$ contain a at position R^p , then they are both facts of I . In our context, however, the core chase matches the UID chase defined above, except at the first round. Thus, modulo the first round, by $\text{Chase}(I, \Sigma_{\text{UID}})$ we refer to the UID chase, which has the Unique Witness Property. See Appendix A for details.

Finite closure. Rosati [16, 18] showed that, while conjunctions of IDs are finitely controllable, even conjunctions of UIDs and FDs may not be. However, Cosmadakis et al. [8] showed how to decide in PTIME the *finite implication* problem for UIDs and FDs: given a conjunction Σ of such dependencies, decide whether a UID or FD is implied by Σ over finite instances. The *finite closure* of Σ is the set of the UIDs and FDs thus implied by Σ in the finite.

Rosati [17] later showed that the finite closure could be used to reduce UQA to FQA for some constraints on relations of arity at most two. Following the same idea, we say that a conjunction of constraints Σ is *finitely controllable up to finite closure* if for every finite instance I , and Boolean CQ q , $(I, \Sigma) \models_{\text{fin}} q$ iff $(I, \Sigma') \models_{\text{unr}} q$, where Σ' is the finite closure of Σ . This implies that we can reduce FQA to UQA, even if finite controllability does not hold.

III. Main Result and Overall Approach

We study open-world query answering for FDs and UIDs. For unrestricted query answering (UQA), the following is already known, from bounds on UQA for UIDs:

Proposition III.1. *UQA for FDs and UIDs has PTIME data complexity and NP-complete combined complexity.*

However, for the *finite case*, even the decidability of FQA for FDs and UIDs is not known. Here is our main result, which is proved in the rest of this paper:

Theorem III.2 (Main theorem). *Conjunctions of FDs and UIDs are finitely controllable up to finite closure.*

From these two results, and an efficient computation of the closure, we deduce that the complexity of FQA matches that of UQA (see Appendix B.3):

Corollary III.3. *FQA for FDs and UIDs has PTIME data complexity and NP-complete combined complexity.*

III.1. Rephrasing with universal models

We prove the main theorem via the notions of *k-sound* and *k-universal instances*.

Definition III.4. For $k \in \mathbb{N}$, we say that a superinstance I of an instance I_0 is *k-sound* for constraints Σ (and for I_0) if for every constant-free CQ q of size $\leq k$ such that $I \models q$, we have $(I_0, \Sigma) \models_{\text{unr}} q$. We say it is *k-universal* if the converse also holds: $I \models q$ whenever $(I_0, \Sigma) \models_{\text{unr}} q$.

The assumption that q is constant-free is without loss of generality: we can always assume that, for each constant $c \in \text{dom}(I_0)$, a fact $P_c(c)$ has been added to I_0 for a fresh unary relation P_c , and c was replaced in q by a existentially quantified variable x_c with the atom $P_c(x_c)$ added to q . So for simplicity we assume from now on that queries are constant-free.

Theorem III.2 is implied by the following (see Appendix B.2):

Theorem III.5 (Universal models). *For every conjunction Σ of FDs Σ_{FD} and UIDs Σ_{UID} closed under finite implication, for every finite instance I_0 that satisfies Σ_{FD} , for any $k \in \mathbb{N}$, there exists a finite superinstance I of I_0 that is k -sound for Σ and satisfies Σ (and hence is k -universal).*

The fact that such an I is k -universal is because any superinstance of I_0 that satisfies Σ must satisfy all CQs q such that $(I_0, \Sigma) \models_{\text{unr}} q$, by definition of \models_{unr} .

We now fix the conjunction Σ of FDs Σ_{FD} and UIDs Σ_{UID} . We assume that Σ is closed under finite implication; in particular, Σ_{FD} and Σ_{UID} in isolation are closed under implication, which implies that Σ_{UID} is transitively closed. We also fix the instance I_0 such that $I_0 \models \Sigma_{\text{FD}}$, and the maximal query size $k \in \mathbb{N}$.

Our goal in the rest of this paper is to construct the finite k -sound superinstance of I_0 that satisfies Σ , thus proving the Universal Models Theorem and hence the Main Theorem.

III.2. Restricting to ACQs, UFDs, and reversible constraints

We first prove the Universal Models Theorem for a restricted class of queries and dependencies, which we now define. We will lift these restrictions later.

First, we define Σ_{UFD} to be the *unary* FDs of Σ_{FD} , and write $\Sigma_{\text{U}} := \Sigma_{\text{UFD}} \wedge \Sigma_{\text{UID}}$. Note that, as we assumed that Σ is closed under finite implication for UFDs and UIDs, the characterization of [8] implies that Σ_{U} also is. We will first construct a k -sound superinstance that only satisfies Σ_{U} ; in Section VII we will show how to adapt the process to also satisfy Σ .

Second, we will first construct a superinstance that is k -sound only for acyclic Boolean queries; in Section VIII we will show how to make the resulting superinstance sufficiently acyclic to be sound for cyclic queries as well.

Hence, in Sections IV, V and VI, we prove the following weakening of the Universal Models Theorem. The restrictions will be lifted in Sections VII and VIII.

Theorem III.6 (Acyclic unary universal models). *There exists a finite superinstance of I_0 that satisfies Σ_{U} and is k -sound for Σ_{U} and ACQ (and hence k -universal for Σ_{U} and ACQ).*

To prove the Acyclic Unary Universal Models Theorem, in Sections IV and V, we will assume the following condition on the structure of the dependencies:

reversible: The following holds about Σ_{U} :

- all UIDs in Σ_{UID} are **reversible** (remember this means that the reverse τ^{-1} of any $\tau \in \Sigma_{\text{UID}}$ is also in Σ_{UID});
- for any positions R^p and R^q occurring in UIDs of Σ_{UID} , if $R^p \rightarrow R^q$ is in Σ_{UFD} then so is $R^q \rightarrow R^p$.

Intuitively, assumption reversible is connected to the finite closure characterization of [8], which adds to Σ_{U} the reverses of any UIDs and UFDs that form a certain cyclic pattern.

Working under assumption reversible, Section IV proves an even weaker version of the Acyclic Unary Universal Models Theorem, which replaces k -soundness by weak-soundness; Section V proves the actual theorem. Assumption reversible is lifted in Section VI to conclude the proof.

IV. Weak-Soundness and Reversible UIDs

The goal of this section is to prove the Acyclic Unary Universal Models Theorem (Theorem III.6) under assumption reversible, replacing k -soundness by *weak-soundness*.

Definition IV.1. A superinstance I' of an instance I is **weakly-sound** if the following holds:

- for any $a \in \text{dom}(I)$ and $R^p \in \text{Pos}(\sigma)$, if $a \in \pi_{R^p}(I')$, then either $a \in \pi_{R^p}(I)$ or $a \in \text{Wants}(I, R^p)$;
- for any $a \in \text{dom}(I') \setminus \text{dom}(I)$ and $R^p, S^q \in \text{Pos}(\sigma)$, if $a \in \pi_{R^p}(I')$ and $a \in \pi_{S^q}(I')$ then $R^p = S^q$ or $R^p \subseteq S^q$ is in Σ_{UID} .

Intuitively, a superinstance is weakly-sound if existing elements were only added to positions where they wanted to appear, and new elements only occur at positions which are connected in Σ_{UID} . This section shows the following:

Proposition IV.2 (Acyclic unary weakly-sound models). *Under assumption reversible, there exists a finite superinstance of I_0 that satisfies Σ_{U} and is weakly-sound.*

The proposition itself will not be reused in the sequel, but the proof introduces some useful concepts to prove the actual Acyclic Unary Universal Models Theorem in Section V.

IV.1. Binary signatures and balanced instances

For simplicity, we first focus on a simplified case with a binary signature, making the following assumption that will be lifted later in this section:

binary: all relations have arity 2 and Σ_{UFD} contains the UFDs $R^1 \rightarrow R^2$ and $R^2 \rightarrow R^1$ for any relation R .

Our approach to construct a weakly-sound superinstance I' of I_0 that satisfies Σ_{U} is then to perform a *completion process* that adds new (binary) facts to connect together elements. As all possible UFDs hold, I' can only contain a new fact $R(a_1, b_2)$ if, for $i \in \{1, 2\}$, $a_i \notin \pi_{R^i}(I_0)$, so that if $a_i \in \text{dom}(I_0)$ then $a_i \in \text{Wants}(I_0, R^i)$ by weak soundness.

One easy situation is when I_0 is *balanced*: for every relation R , we can construct a bijection between the elements that want to be in R^1 and those that want to be in R^2 :

Definition IV.3. An instance I is **balanced** if, for every two positions R^p and R^q such that $R^p \rightarrow R^q$ and $R^q \rightarrow R^p$ are in Σ_{UFD} , we have $|\text{Wants}(I, R^p)| = |\text{Wants}(I, R^q)|$.

If I_0 is balanced, we can show the Acyclic Unary Weakly-Sound Models Proposition under assumption binary, simply by pairing together elements, without adding any new ones:

Proposition IV.4. Assuming binary and reversible, any balanced finite instance I satisfying Σ_{UFD} has a finite weakly-sound superinstance I' that satisfies Σ_{U} , with $\text{dom}(I') = \text{dom}(I)$.

However, our instance I_0 may not be balanced. The idea is then to balance it by adding “helper” elements and assigning them to positions, as the following example shows:

Example IV.5. Consider three binary relations R, S, T , with the UIDs $R^2 \subseteq S^1$, $S^2 \subseteq T^1$, $T^2 \subseteq R^1$ and their reverses, and the FDs prescribed by assumption binary. Consider $I_0 := \{R(a, b)\}$. We have $a \in \text{Wants}(I_0, T^2)$ and $b \in \text{Wants}(I_0, S^1)$; however $\text{Wants}(I_0, S^2) = \text{Wants}(I_0, T^1) = \emptyset$, so I_0 is not balanced.

Still, we can construct the weakly-sound superinstance $I := \{R(a, b), S(b, c), T(c, a)\}$ that satisfies the constraints. Intuitively, we have added a “helper” element c and “assigned” it to the positions S^1 and T^2 , which are connected by the UIDs.

We now formalize this idea of constructing weakly-sound superinstances where the domain is augmented with *helper elements*. We first need to understand at which positions the helpers can appear to avoid violating weak-soundness:

Definition IV.6. For any two positions R^p and S^q , we write $R^p \sim_{\text{ID}} S^q$ when $R^p = S^q$ or when $R^p \subseteq S^q$, and hence $S^q \subseteq R^p$ by assumption reversible, are in Σ_{UID} .

As Σ_{UID} is transitively closed, \sim_{ID} is an equivalence relation. Our idea to construct weakly-sound superinstances is thus to first decide on the helpers that we want to add, and the \sim_{ID} -class to which we want to assign them, following the definition of weak-soundness. We represent this choice as a *partially-specified superinstance*, or *pssinstance*:

Definition IV.7. A *pssinstance* of an instance I is a triple $P = (I, \mathcal{H}, \lambda)$ where \mathcal{H} is a finite set of **helpers** and λ maps each $h \in \mathcal{H}$ to an \sim_{ID} -class $\lambda(h)$.

We define $\text{Wants}(P, R^p) := \text{Wants}(I, R^p) \sqcup \{h \in \mathcal{H} \mid R^p \in \lambda(h)\}$. This allows us to talk of P being *balanced* following Definition IV.3.

A superinstance I' of I is a **realization** of P if $\text{dom}(I') = \text{dom}(I) \sqcup \mathcal{H}$, and, for any fact $R(a)$ of $I' \setminus I$ and $R^p \in \text{Pos}(R)$, we have $a_p \in \text{Wants}(P, R^p)$.

Example IV.8. In Example IV.5, a pssinstance of I_0 is $P := (I_0, \{c\}, \lambda)$ where $\lambda(c) := \{S^1, T^2\}$, and I is a realization of P .

It is always possible to balance an instance by adding helpers:

Lemma IV.9 (Balancing). For any finite instance I , if I satisfies Σ_{UFD} then it has a balanced pssinstance.

From there, we can construct realizations like we constructed superinstances in Lemma IV.4.

Lemma IV.10 (Binary realizations). For any balanced pssinstance P of an instance I that satisfies Σ_{UFD} , we can construct a realization of P that satisfies Σ_{U} .

We then observe that realizations are weakly-sound superinstances of I_0 .

Lemma IV.11 (Binary realizations are completions). If I' is a realization of a pssinstance of I then it is a weakly-sound superinstance of I .

We have thus proved the Acyclic Unary Weakly-Sound Models Proposition under assumptions binary and reversible, using the completion process formed by combining the three above lemmas.

IV.2. Arbitrary arity and piecewise realizations

We now lift assumption binary (but retain assumption reversible). We show how to generalize the previous constructions to the arbitrary arity case. Contrary to the binary situation, we will see later that the resulting completion process needs to assume that a certain *saturation* process has been applied to I_0 beforehand.

The definition of balanced instances (Definition IV.3) generalizes to arbitrary arity, and we can show that the Balancing Lemma (Lemma IV.9) still holds. We keep the definition of pssinstance (Definition IV.7) but need to change the notion of realization. We replace it by *piecewise realizations*, which are defined on subsets of positions that are connected in Σ_{UFD} .

Definition IV.12. For any two positions R^p and R^q , we write $R^p \leftrightarrow_{\text{FUN}} R^q$ whenever $R^p \rightarrow R^q$ and $R^q \rightarrow R^p$ are in Σ_{UFD} .

By transitivity of Σ_{UFD} , $\leftrightarrow_{\text{FUN}}$ is clearly an equivalence relation. We number the $\leftrightarrow_{\text{FUN}}$ -classes of $\text{Pos}(\sigma)$ as Π_1, \dots, Π_n and define *piecewise instances* by their projections to the Π_i :

Definition IV.13. A *piecewise instance* is an n -tuple $PI = (K_1, \dots, K_n)$, where each K_i is a set of $|\Pi_i|$ -tuples, indexed by Π_i for convenience. The **domain** of PI is $\text{dom}(PI) := \bigcup_i \text{dom}(K_i)$. For $1 \leq i \leq n$ and $R^p \in \Pi_i$, we write $\pi_{R^p}(PI) := \pi_{R^p}(K_i)$.

We use this to define *piecewise realizations* of pssinstances:

Definition IV.14. A piecewise instance $PI = (K_1, \dots, K_n)$ is a **piecewise realization** of the pssinstance $P = (I, \mathcal{H}, \lambda)$ if:

- $\pi_{\Pi_i}(I) \subseteq K_i$ for all $1 \leq i \leq n$,
- $\text{dom}(PI) = \text{dom}(I) \sqcup \mathcal{H}$,
- for all $1 \leq i \leq n$, for all $R^p \in \Pi_i$, for every tuple $\mathbf{a} \in K_i \setminus \pi_{\Pi_i}(I)$, we have $a_p \in \text{Wants}(P, R^p)$.

In order to generalize the Binary Realizations Lemma (Lemma IV.10), we need to talk of a piecewise instance PI “satisfying” Σ_U . For Σ_{UFD} , we require that PI respects the UFDs within each $\leftrightarrow_{\text{FUN}}$ -class. For Σ_{UID} , we define it directly from the projections of PI .

Definition IV.15. A piecewise instance PI is Σ_{UFD} -**compliant** if, for all $1 \leq i \leq n$, there are no two tuples $\mathbf{a} \neq \mathbf{b}$ in K_i such that $a_p = b_p$ for some $R^p \in \Pi_i$.

PI is Σ_{UID} -**compliant** if $\text{Wants}(PI, \tau) := \pi_{R^p}(PI) \setminus \pi_{S^q}(PI)$ is empty for all $\tau \in \Sigma_{\text{UID}}$.

PI is Σ_U -**compliant** if it is Σ_{UFD} - and Σ_{UID} -compliant.

We can then generalize the Binary Realizations Lemma:

Lemma IV.16 (Realizations). For any balanced pssinstance P of an instance I that satisfies Σ_{UFD} , we can construct a Σ_U -compliant piecewise realization of P .

Example IV.17. Consider a 4-ary relation R and the UFDs $\tau : R^1 \subseteq R^2$, $\tau' : R^3 \subseteq R^4$ and their reverses, and the UFDs $\phi : R^1 \rightarrow R^2$, $\phi' : R^3 \rightarrow R^4$ and their reverses. We have $\Pi_1 = \{R^1, R^2\}$ and $\Pi_2 = \{R^3, R^4\}$. Consider $I_0 := \{R(a, b, c, d)\}$, which is balanced, and the balanced pssinstance $P := (I_0, \emptyset, \lambda)$, where λ is the empty function. A Σ_U -compliant piecewise realization of P is $PI := (\{(a, b), (b, a)\}, \{(c, d), (d, c)\})$.

We now transform the Σ_U -compliant piecewise realization PI into a weakly-sound superinstance, generalizing the “Binary Realizations Are Completions” Lemma (Lemma IV.11), and completing the description of our completion process. The idea is to expand each tuple \mathbf{t} of each K_i to an entire fact F_t of the corresponding relation.

However, to fill the other positions of F_t , we will need to reuse existing elements of I_0 . For this, we want I_0 to contain some R -fact for every relation R that occurs in $\text{Chase}(I_0, \Sigma_{\text{UID}})$.

Definition IV.18. A relation R is **achieved** (by I and Σ_{UID}) if there is some R -fact in $\text{Chase}(I, \Sigma_{\text{UID}})$.

A superinstance I' of an instance I is **relation-saturated** (for Σ_{UID}) if every achieved relation (by I and Σ_{UID}) occurs in I' .

Example IV.19. Consider two binary relations R and T and a unary relation S , the UFDs $\tau : S^1 \subseteq R^1$, $\tau' : R^2 \subseteq T^1$ and their reverses, no UFDs, and the non-relation-saturated instance $I_0 := \{S(a)\}$ which is trivially balanced.

$P := (I_0, \emptyset, \lambda)$, with λ the empty function, is a pssinstance of I , and $PI := (\{(a)\}, \emptyset, \{(a)\}, \emptyset, \emptyset)$, where Π_1 and Π_3 are the $\leftrightarrow_{\text{FUN}}$ -classes of R^1 and S^1 , is a Σ_U -compliant piecewise realization of P . However, we cannot easily complete PI to a superinstance of I_0 satisfying τ and τ' , because, to create the fact $R(a, \bullet)$, we need to create an element to fill position R^2 , and this would introduce a violation of τ' . Intuitively, this is because I_0 is not relation-saturated.

Consider instead the instance $I_1 := I_0 \sqcup \{S(c), R(c, d), T(d)\}$. We can complete I_1 to satisfy τ and τ' by adding the fact $R(a, d)$, reusing the element d to fill position R^2 .

Clearly, initial chasing on I_0 ensures relation-saturation:

Lemma IV.20 (Relation-saturated solutions). *The result of performing sufficiently many chase rounds on any instance I is relation-saturated.*

Relation-saturation ensures that we can reuse existing elements when completing PI . This allows us to perform the last step of the completion process:

Lemma IV.21 (Using realizations to get completions). *For any finite relation-saturated instance I that satisfies Σ_{UFD} , from a Σ_{U} -compliant piecewise realization PI of a pssinstance of I , we can construct a finite weakly-sound superinstance of I that satisfies Σ_{U} .*

We can now prove the Acyclic Unary Weakly-Sound Models Proposition. Consider our initial finite instance I_0 , that satisfies Σ_{UFD} , and chase it to a finite relation-saturated superinstance I'_0 using the Relation-Saturated Solutions Lemma. By the Unique Witness Property, I'_0 still satisfies Σ_{UFD} , and it is clearly a weakly-sound superinstance of I_0 .

Now, perform the completion process: construct a balanced pssinstance P of I'_0 using the Balancing Lemma (Lemma IV.9), and a finite Σ_{U} -compliant piecewise realization PI of P by the Realizations Lemma (Lemma IV.16). Then, use the realization PI with Lemma IV.21 to construct the finite weakly-sound superinstance I of I'_0 that satisfies Σ_{U} . I is clearly also a weakly-sound superinstance of I_0 , so the result is proven.

V. k -Soundness and Reversible UIDs

We now move from weak-soundness to k -soundness, to prove the Acyclic Unary Universal Models Theorem (Theorem III.6), still making assumption reversible.

We first introduce the notion of *aligned superinstances* that we use to maintain k -soundness, and give the saturation process that generalizes relation-saturation. We then define a notion of *thrifty chase steps*, and a completion process that uses these chase steps to repair UID violations in the instance.

V.1. Aligned superinstances and fact-saturation

We ensure k -soundness by maintaining a k -bounded simulation from our superinstance of I_0 to the chase $\text{Chase}(I_0, \Sigma_{\text{UID}})$. Indeed, $\text{Chase}(I_0, \Sigma_{\text{UID}})$ is a universal model for Σ_{UID} , and it satisfies Σ_{FD} (by the Unique Witness Property, and because I_0 does). Hence, it is in particular k -sound for Σ . Now, as acyclic queries of size $\leq k$ are preserved through k -bounded simulations, superinstances of I_0 with a k -bounded simulation to $\text{Chase}(I_0, \Sigma_{\text{UID}})$ are indeed k -sound for ACQ.

Definition V.1. For I, I' two instances, $a \in \text{dom}(I)$, $b \in \text{dom}(I')$, and $n \in \mathbb{N}$, we write $(I, a) \leq_n (I', b)$ if, for any fact $R(a)$ of I with $a_p = a$ for some $R^p \in \text{Pos}(R)$, there exists a fact $R(b)$ of I' such that $b_p = b$, and $(I, a_q) \leq_{n-1} (I', b_q)$ for all $R^q \in \text{Pos}(R)$. The base case $(I, a) \leq_0 (I', b)$ always holds.

An n -bounded simulation from I to I' is a mapping sim such that for all $a \in \text{dom}(I)$, $(I, a) \leq_n (I', \text{sim}(a))$.

We write $a \simeq_n b$ for $a, b \in \text{dom}(I)$ if both $(I, a) \leq_n (I, b)$ and $(I, b) \leq_n (I, a)$; this is an equivalence relation on $\text{dom}(I)$.

Lemma V.2. For any instance I and ACQ q of size $\leq n$ such that $I \models q$, if there is an n -bounded simulation from I to I' , then $I' \models q$.

We accordingly give a name to superinstances of I_0 that have a k -bounded simulation to the chase. For convenience, we also require them to be finite and satisfy Σ_{UFD} . For technical reasons we require that the simulation is the identity on I_0 , that it does not map other elements to I_0 , and that elements occur in the superinstance at least at the position where their sim-image was introduced in the chase:

Definition V.3. An *aligned superinstance* $J = (I, \text{sim})$ of I_0 is a finite superinstance I of I_0 that satisfies Σ_{UFD} , and a k -bounded simulation sim from I to $\text{Chase}(I_0, \Sigma_{\text{UID}})$ such that $\text{sim}|_{I_0}$ is the identity and $\text{sim}|_{(I \setminus I_0)}$ maps to $\text{Chase}(I_0, \Sigma_{\text{UID}}) \setminus I_0$.

Further, for any $a \in \text{dom}(I) \setminus \text{dom}(I_0)$, letting R^p be the position where $\text{sim}(a)$ was introduced in $\text{Chase}(I_0, \Sigma_{\text{UID}})$, we require that $a \in \pi_{R^p}(I)$.

Before we perform the *completion process* that allows us to satisfy Σ_{UID} , we need to perform a *saturation process*, like relation-saturation in the previous section. Instead of achieving all relations, we want the aligned superinstance to achieve all *fact classes*:

Definition V.4. A *fact class* is a pair (R^p, C) of a position $R^p \in \text{Pos}(\sigma)$ and a $|R|$ -tuple of \simeq_k -classes of elements of $\text{Chase}(I_0, \Sigma_{\text{UID}})$. The dependency on k is omitted for brevity.

The *fact class* of a fact $F = R(\mathbf{a})$ of $\text{Chase}(I_0, \Sigma_{\text{UID}}) \setminus I_0$ is (R^p, C) , where a_p is the exported element of F and C_i is the \simeq_k -class of a_i in $\text{Chase}(I_0, \Sigma_{\text{UID}})$ for all $R^i \in \text{Pos}(R)$.

A fact class (R^p, C) is *achieved* if it is the fact class of some fact of $\text{Chase}(I_0, \Sigma_{\text{UID}}) \setminus I_0$. We write AFactCl for the set of all achieved fact classes (for brevity, the dependence on I_0 , Σ_{UID} , and k is omitted from notation).

An aligned superinstance $J = (I, \text{sim})$ is *fact-saturated* if, for any achieved fact class $D = (R^p, C)$ in AFactCl , there is a fact $F_D = R(\mathbf{a})$ of $I \setminus I_0$ such that $\text{sim}(a_i) \in C_i$ for all $R^i \in \text{Pos}(R)$. We say that F_D *achieves* D in J .

Lemma V.5. For any initial instance I_0 , set Σ_{UID} of UIDs, and $k \in \mathbb{N}$, AFactCl is finite.

We now define our saturation process: chase I_0 until all fact classes are achieved, which is possible in finitely many rounds thanks to the above lemma. The result is easily seen to be a fact-saturated aligned superinstance:

Lemma V.6 (Fact-saturated solutions). The result I of performing sufficiently many chase rounds on I_0 is such that $J_0 = (I, \text{id})$ is a fact-saturated aligned superinstance of I_0 .

We thus obtain a fact-saturated aligned superinstance J_0 of I_0 , which we now want to complete to one that satisfies Σ_{UID} .

V.2. Fact-thrifty completion

Our general method to repair UID violations in J_0 is to apply a form of chase step on aligned superinstances, which may reuse elements: *thrifty chase steps*. To define them, we first distinguish *dangerous* and *non-dangerous* positions, which determine how we may reuse elements when chasing.

Definition V.7. We say a position $S^r \in \text{Pos}(\sigma)$ is *dangerous* for a position $S^q \neq S^r$ if $S^r \rightarrow S^q$ is in Σ_{UFD} , and write $S^r \in \text{Dng}(S^q)$. Otherwise, S^r is *non-dangerous*, written $S^r \in \text{NDng}(S^q)$. Note that $\{S^q\} \sqcup \text{Dng}(S^q) \sqcup \text{NDng}(S^q) = \text{Pos}(S)$.

Definition V.8 (Thrifty chase steps). Let $J = (I, \text{sim})$ be an aligned superinstance of I_0 , let $\tau : R^p \subseteq S^q$ be a UID of Σ_{UID} , and let $F_a = R(\mathbf{a})$ be an active fact for τ in I .

Because sim is a 1-bounded simulation, $\text{sim}(a_p) \in \pi_{Rp}(\text{Chase}(I_0, \Sigma_{\text{UID}}))$, so, because the chase satisfies τ , there is a fact $F_w = S(b')$ in $\text{Chase}(I_0, \Sigma_{\text{UID}})$ with $b'_q = \text{sim}(a_p)$; we call F_w the **chase witness**.

Applying a **thrifty chase step** on F_a for τ yields an aligned superinstance $J' = (I', \text{sim}')$. We define I' as I plus a new fact $F_n = S(b)$, where $b_q = a_p$ and the b_r for $S^r \neq S^q$ may be elements of $\text{dom}(J)$ or fresh elements. We require that:

- for $S^r \in \text{NDng}(S^q)$, $b_r \in \pi_{Sr}(J)$ (so they are not fresh)
- for $S^r \in \text{Dng}(S^q)$, $b_r \notin \pi_{Sr}(J)$ (so they may be fresh)
- for $S^r \neq S^q$, if b_r is not fresh then $\text{sim}(b_r) \simeq_k b'_r$.

We define sim' by extending sim to $\text{dom}(J')$: we set $\text{sim}'(b_r) := b'_r$ whenever b_r is fresh.

A **fact-thrifty chase step** is a thrifty chase step where we choose one fact $F_t = S(c)$ of $J \setminus I_0$ that achieves the fact class of F_w (that is, $\text{sim}(c_i) \simeq_k b'_i$ for all i), and use F_t to define $b_r := c_r$ for all $S^r \in \text{NDng}(S^q)$.

The chase step is **fresh** if b_r is fresh for all $S^r \in \text{Dng}(S^q)$.

Thrifty chase steps may in general violate Σ_{UFD} , but fact-thrifty chase steps never do. For this reason, we will only use fact-thrifty chase steps in this section. The point of working with fact-saturated aligned superinstances is that we can ensure that a suitable F_t always exists. We thus claim:

Lemma V.9 (Fact-thrifty chase steps). *For any fact-saturated aligned superinstance J , the result J' of a fact-thrifty chase step on J is indeed a well-defined aligned superinstance where the former active fact F_a is no longer active.*

We now claim that we can expand fact-saturated superinstances to satisfy Σ_{UID} , using fact-thrifty chase steps:

Proposition V.10 (Fact-thrifty completion). *Under assumption reversible, for any fact-saturated aligned superinstance J of I_0 , we can expand J by fact-thrifty chase steps to a fact-saturated aligned superinstance J' of I_0 that satisfies Σ_{UID} .*

This proposition allows us to prove the Acyclic Unary Universal Models Theorem (Theorem III.6) under assumption reversible. Indeed, consider the fact-saturated aligned superinstance J_0 produced by the Fact-Saturated Solutions Lemma (Lemma V.6). Applying the Fact-Thrifty Completion Proposition to J_0 yields a fact-saturated aligned superinstance J' , which is a finite k -sound superinstance of I_0 that satisfies Σ_{UFD} and satisfies Σ_{UID} .

The rest of this section sketches the proof of the Proposition (see Appendix D.5 for the full proof). The idea is to construct, as in Section IV, a balanced pssinstance P of the input aligned superinstance J , and a Σ_{U} -compliant piecewise realization PI of P . Now, instead of completing the facts of PI to add them directly to J , we add them one by one, using fact-thrifty chase steps, to ensure that alignedness is preserved.

The only problematic point is that PI could connect together elements that have dissimilar sim-images, violating alignedness. However, we show that, up to chasing for $k + 1$ rounds on the initial J with fresh fact-thrifty chase steps before constructing P , we can ensure what we call *k-reversibility*: all elements that want to be at some position R^p in J have a sim-image whose \simeq_k -class only depends on R^p . Once we have ensured this, we can essentially stop worrying about sim-images, because respecting weak-soundness, as PI does, is sufficient.

The reason why $k + 1$ chasing rounds suffice to ensure this is by a general structural observation on the UID chase: when the last k UIDs applied to an element a of $\text{Chase}(I_0, \Sigma_{\text{UID}})$ are reversible (as is the case here, by assumption reversible), the \simeq_k -class of a only depends on the \sim_{ID} -class of the position where it was introduced, and not on its exact history. Formally:

Theorem V.11 (Chase locality theorem). *For any instance I_0 , transitively closed set of UIDs Σ_{UID} , and $n \in \mathbb{N}$, for any two elements a and b respectively introduced at positions R^p and S^q in $\text{Chase}(I_0, \Sigma_{\text{UID}})$ such that $R^p \sim_{\text{ID}} S^q$, if the last n UIDs applied to create a and b are reversible, then $a \simeq_n b$.*

VI. Arbitrary UIDs: Lifting Assumption reversible

This section concludes the proof of the Acyclic Unary Universal Models Theorem (Theorem III.6) by removing assumption reversible. We do so by splitting Σ_{UID} in subsets that can be satisfied sequentially:

Definition VI.1. *For any $\tau, \tau' \in \Sigma_{\text{UID}}$, we write $\tau \rightarrow \tau'$ when we can write $\tau = R^p \subseteq S^q$ and $\tau' = S^r \subseteq T^u$ with $S^q \neq S^r$, and the UFD $S^r \rightarrow S^q$ is in Σ_{UFD} . An **ordered partition** (P_1, \dots, P_n) of Σ_{UID} is a partition of Σ_{UID} (i.e., $\Sigma_{\text{UID}} = \bigsqcup_i P_i$) such that for any $\tau \in P_i, \tau' \in P_j$, if $\tau \rightarrow \tau'$ then $i \leq j$.*

The notion of ordered partition is useful because thrifty chase steps can only cause new UID violations at the dangerous positions of the new fact. This implies the following:

Lemma VI.2. *Let J be an aligned superinstance of I_0 and J' be the result of applying a thrifty chase step on J for a UID τ of Σ_{UID} . Assume that a UID τ' of Σ_{UID} was satisfied by J but is not satisfied by J' . Then $\tau \rightarrow \tau'$.*

Hence, given an ordered partition of Σ_{UID} , once we have satisfied the UIDs of the first i classes P_1, \dots, P_i , then this property is preserved while we do thrifty chasing with $P_j, j > i$. So if we can satisfy each P_i individually with thrifty chase steps, then we can satisfy Σ_{UID} by satisfying P_1, \dots, P_n .

Of course, the point of partitioning Σ_{UID} is to be able to control the structure of the UIDs in each class:

Definition VI.3. *We call $P \subseteq \Sigma_{\text{UID}}$ **reversible** if it is transitively closed (as Σ_{UID} is) and satisfies assumption reversible.*

*We say $P \subseteq \Sigma_{\text{UID}}$ is **trivial** if we have $P = \{\tau\}$ for some $\tau \in \Sigma_{\text{UID}}$ such that $\tau \not\rightarrow \tau$. An ordered partition is **manageable** if all of its classes are either reversible or trivial.*

If $P \subseteq \Sigma_{\text{UID}}$ is reversible, then the previous section describes how to complete with thrifty chase steps any fact-saturated aligned superinstance of I_0 to one that satisfies P . If P is trivial, it follows directly from Lemma VI.2 that we can satisfy it:

Corollary VI.4. *For any trivial class $\{\tau\}$, performing one chase round on an aligned fact-saturated superinstance J of I_0 by fresh fact-thrifty chase steps for τ yields an aligned superinstance J' of I_0 that satisfies τ .*

We now claim that we can construct a manageable partition of Σ_{UID} . We build it as a topological sort of the strongly connected components (SCCs) of the directed graph on Σ_{UID} defined by \rightarrow , with the technical complication that SCCs must be closed under UID reversal. The construction relies on the fact that Σ_{UID} is closed under finite implication, as characterized by Cosmadakis et al. [8].

Lemma VI.5. *Any conjunction Σ_{UID} of UIDs closed under finite implication has a manageable partition.*

Example VI.6. *Consider the UIDs $\tau_R : R^1 \subseteq R^2$, $\tau_S : S^1 \subseteq S^2$, $\tau : R^3 \subseteq S^3$, and the UFDs $\phi_R : R^1 \rightarrow R^2$, $\phi_S : S^1 \rightarrow S^2$, $\phi'_R : R^3 \rightarrow R^1$, and $\phi'_S : S^1 \rightarrow S^3$. The UIDs τ_R^{-1} and τ_S^{-1} , and UFDs ϕ_R^{-1} , ϕ_S^{-1} , and $R^3 \rightarrow R^2$, $S^2 \rightarrow S^3$, are finitely implied. A manageable partition is $(\{\tau_R, \tau_R^{-1}\}, \{\tau\}, \{\tau_S, \tau_S^{-1}\})$, where the first and third classes are reversible and the second is trivial.*

We can now conclude the proof of the Acyclic Unary Universal Models Theorem (Theorem III.6). We first note that the Fact-Saturated Solutions Lemma (Lemma V.6) does not use assumption reversible, so we apply it (with Σ_{UID}) to obtain from I_0 an aligned fact-saturated superinstance J_1 of I_0 . This is the **saturation process**.

We now satisfy Σ_{UID} by a **completion process**. Build a manageable partition (P_1, \dots, P_n) of Σ_{UID} , by Lemma VI.5. Now, for $1 \leq i \leq n$, use fact-thrifty chase steps by UIDs of P_i to extend the fact-saturated aligned superinstance J_i to a larger one J_{i+1} that satisfies P_i . If P_i is trivial, use Corollary VI.4. If P_i is reversible, apply the Fact-Thrifty Completion Proposition (Proposition V.10), taking Σ_{UID} to be P_i . By Lemma VI.2, the result J_{i+1} satisfies $\bigcup_{j \leq i} P_j$.

Hence the result J_{n+1} of the completion process is an aligned superinstance of I_0 that satisfies Σ_{UID} ; as an aligned superinstance, it is also finite, satisfies Σ_{UFD} , and is k -sound for ACQ; so it is k -universal for Σ_{U} and ACQ. This concludes the proof of the Acyclic Unary Universal Models Theorem.

VII. Higher-Arity FDs

We now bootstrap the Acyclic Unary Universal Models Theorem (Theorem III.6) to the Universal Models Theorem (Theorem III.5). The first step is to change our construction to avoid violating higher-arity FDs, namely, show the following, which applies to $\Sigma = \Sigma_{\text{UID}} \wedge \Sigma_{\text{FD}}$ rather than $\Sigma_{\text{U}} = \Sigma_{\text{UID}} \wedge \Sigma_{\text{UFD}}$:

Theorem VII.1 (Acyclic universal models). *There is a finite superinstance of I_0 that is k -universal for Σ and ACQ queries.*

The problem to address is that our completion process to satisfy Σ_{UID} was defined with fact-thrifty chase steps, which reuse elements from the same facts at the same positions multiple times. This may violate Σ_{FD} , and we can show that is the only point where we do so in the construction.

The goal of this section is to define a new version of thrifty chase steps that preserves Σ_{FD} rather than just Σ_{UFD} ; we call them *envelope-thrifty chase steps*. We first describe the new saturation process designed for them. Second, we define how they work, redefine the completion process of the previous section to use them, and use this new completion process to prove the Acyclic Universal Models Theorem above.

VII.1. Envelopes and saturation

We start by defining a new notion of saturated instances. Recall the notions of fact classes (Definition V.4) and thrifty chase steps (Definition V.8). When a thrifty chase step wants to create a fact F_n whose chase witness F_w has fact class (R^p, C) , it needs elements to reuse in F_n at positions of $\text{NDng}(R^p)$. They must have the right sim-image and must already occur at the positions where they are reused.

Fact-thrifty chase steps reuse a tuple of elements from one fact F_r , and thus apply to *fact-saturated instances* with one fact for each class. Our new notion of envelope-thrifty chase steps will need saturated instances that have *multiple* reusable tuples. A set of such tuples is called an *envelope* for (R^p, C) :

Definition VII.2. Consider $D = (R^p, C)$ in AFactCl , and write $O := \text{NDng}(R^p)$. An *envelope* E for D and for an aligned superinstance $J = (I, \text{sim})$ of I_0 is a non-empty set of $|O|$ -tuples indexed by O , with domain $\text{dom}(I)$, such that:

- for every FD $\phi : R^L \rightarrow R^r$ of Σ_{FD} with $R^L \subseteq O$ and $R^r \in O$, E satisfies ϕ (seeing its tuples as facts on O);
- for every FD $\phi : R^L \rightarrow R^r$ of Σ_{FD} with $R^L \subseteq O$ and $R^r \notin O$, for all $\mathbf{t}, \mathbf{t}' \in E$, $\pi_{R^L}(\mathbf{t}) = \pi_{R^L}(\mathbf{t}')$ implies $\mathbf{t} = \mathbf{t}'$;

- for every $a \in \text{dom}(E)$, there is exactly one position $R^q \in O$ such that $a \in \pi_{R^q}(E)$; and then we also have $a \in \pi_{R^q}(J)$;
- for any fact $F = R(a)$ of J and $R^q \in O$, if $a_q \in \pi_{R^q}(E)$, then F achieves D in J and $\pi_O(a) \in E$.

Intuitively, the tuples in the envelope E satisfy the UFDs of Σ_{UFD} within $\text{NDng}(R^p)$, and never overlap on positions that determine a position out of $\text{NDng}(R^p)$. Further, their elements already occur at the positions where they will be reused, and have the right sim-image for the fact class D . To simplify the reasoning, we also impose that each element of E is used at only one position, and occurs at that position only in facts which achieve D and whose projection to $\text{NDng}(R^p)$ is in E .

Depending on O , it may be possible to use a singleton tuple as the envelope, like fact-thrifty chase steps, and not violate Σ_{FD} . The class is then *safe*. Otherwise, we focus on the envelope tuples which do not appear in the instance yet.

Definition VII.3. We call (R^p, C) in AFactCl *safe* if there is no FD $R^L \rightarrow R^r$ in Σ_{FD} with $R^L \subseteq \text{NDng}(R^p)$ and $R^r \not\subseteq \text{NDng}(R^p)$.

Letting E be an envelope for (R^p, C) and J be an aligned superinstance, the **remaining tuples** of E are $E \setminus \pi_{\text{NDng}(R^p)}(J)$ if (R^p, C) is unsafe, and E if it is safe.

We now introduce the notion of *global envelopes*, that give us one envelope per class of AFactCl . This leads to our new notion of saturation: a saturated instance has a global envelope with many remaining tuples in the unsafe classes. Note that this implies fact-saturation.

Definition VII.4. A *global envelope* \mathcal{E} for an aligned superinstance $J = (I, \text{sim})$ of I_0 is a mapping from each $D \in \text{AFactCl}$ to an envelope $\mathcal{E}(D)$ for D and J , such that the envelopes have pairwise disjoint domains.

We call J *n-envelope-saturated* if it has a global envelope \mathcal{E} such that $\mathcal{E}(D)$ has $\geq n$ remaining tuples for all unsafe $D \in \text{AFactCl}$. J is *envelope-saturated* if it is n -envelope-saturated for $n > 0$, and *envelope-exhausted* otherwise.

We now justify that we can make arbitrarily saturated superinstances of I_0 (the switch to I'_0 is a technicality):

Proposition VII.5 (Sufficiently envelope-saturated solutions). For any $K \in \mathbb{N}$ and instance I_0 , we can build a superinstance I'_0 of I_0 that is k -sound for CQ, and an aligned superinstance J of I'_0 that satisfies Σ_{FD} and is $(K|J|)$ -envelope-saturated.

Example VII.6. For simplicity, we work with instances rather than aligned superinstances. Consider $I_0 := \{S(a), T(z)\}$, the UFDs $\tau : S^1 \subseteq R^1$ and $\tau' : T^1 \subseteq R^1$ for a 3-ary relation R , and the FD $\phi : R^2 R^3 \rightarrow R^1$. Consider $I := I_0 \sqcup \{R(a, b, c)\}$ obtained by one chase step of τ on $S(a)$. It would violate ϕ to perform a fact-thrifty chase step of τ' on z to create $R(z, b, c)$, reusing (b, c) at $\text{NDng}(R^1) = \{R^2, R^3\}$.

Now, consider the k -sound $I'_0 := \{S(a), T(z), S(a'), S(z')\}$, and $I' := I'_0 \sqcup \{R(a, b, c), R(a', b', c')\}$ obtained by two chase steps. The two facts $R(a, b, c)$ and $R(a', b', c')$ would be mapped to the same fact class D , so we can define $E(D) := \{(b, c), (b', c'), (b', c), (b, c')\}$. We can now satisfy Σ_{UFD} on I' without violating ϕ , with two envelope-thrifty chase steps that reuse the remaining tuples (b', c) and (b, c') of $E(D)$.

The crucial result needed for the Sufficiently Envelope-Saturated Proposition is the following, which may be of independent interest, and is proved in Appendix F.2 using a combinatorial construction. The fact that unary keys are problematic is the reason why we handle safe classes differently.

Theorem VII.7 (Dense interpretations). *For any set Σ_{FD} of FDs over a relation R with no unary key, and $K \in \mathbb{N}$, there exists a non-empty instance I of R that satisfies Σ_{FD} and has at least $K |\text{dom}(I)|$ facts.*

Hence, we have defined the new notion of n -envelope-saturation, and a saturation process to achieve it: the Sufficiently Envelope-Saturated Solutions Proposition. Unlike the Fact-Saturated Solutions Lemma, where one fact of each class was enough, we have shown that envelope-saturated superinstances may have an arbitrarily high saturation relative to the instance size.

VII.2. Envelope-thrifty chase steps

We can now introduce *envelope-thrifty chase steps*:

Definition VII.8. *Envelope-thrifty chase steps are thrifty chase steps (Definition V.8) applicable to envelope-saturated aligned superinstances. Let S^q be the exported position of the new fact F_n , let $F_w = S(\mathbf{b}')$ be the chase witness, and let $D = (S^q, \mathbf{C}) \in \text{AFactCl}$ be the fact class of F_w . We choose some remaining tuple \mathbf{t} of $\mathcal{E}(D)$ and define $b_r := t_r$ for all $S^r \in \text{NDng}(S^q)$.*

Recall from Lemma V.9 that fact-thrifty chase steps apply to fact-saturated aligned superinstances, and never violate Σ_{UFD} . Similarly, envelope-thrifty chase steps apply to envelope-saturated aligned superinstances, and never violate Σ_{FD} :

Lemma VII.9. *For $n > 0$, for any n -envelope-saturated aligned superinstance J that satisfies Σ_{FD} , the result J' of an envelope-thrifty chase step on J is an $(n - 1)$ -envelope-saturated superinstance that satisfies Σ_{FD} .*

We now modify the Fact-Thrifty Completion Proposition (Proposition V.10), generalized without assumption reversible as in the previous section, to use envelope-thrifty chase steps instead of fact-thrifty chase steps. This is possible because the choice of reused elements at non-dangerous positions makes no difference in terms of applicable UIDs, as they already occur at the position where they are reused. Hence, we can perform the exact same process as before (except the non-dangerous reuses), using Lemma VII.9 to justify that Σ_{FD} is preserved; but we must abort if we reach an envelope-exhausted instance:

Proposition VII.10 (Envelope-thrifty completion). *For any envelope-saturated aligned superinstance J of I_0 that satisfies Σ_{FD} , we can obtain by envelope-thrifty chase steps an aligned superinstance J' of I_0 , such that J' is either envelope-exhausted or satisfies Σ .*

The last problem to address is exhaustion. Unlike fact-saturation, envelope-saturation “runs out”; whenever we use a remaining tuple \mathbf{t} in a chase step to create F_n and obtain a new aligned superinstance J' , then we cannot use \mathbf{t} again in J' . So we must start with a sufficiently envelope-saturated superinstance, and we must control how many chase steps are applied in the envelope-thrifty completion process. From the details of our construction, we can show the following:

Lemma VII.11 (Envelope blowup). *There exists $B \in \mathbb{N}$ depending only on k and Σ_{U} such that, for any aligned superinstance $J = (I, \text{sim})$ of I_0 , and global envelope \mathcal{E} , letting $J' = (I', \text{sim}')$ be the result of the envelope-thrifty completion process, we have $|I'| < B |I|$.*

We can now conclude the proof of the Acyclic Universal Models Theorem (Theorem III.6) that we stated at the beginning of this section. Start by applying the saturation process of the Sufficiently Envelope-Saturated Solutions Proposition to obtain an aligned superinstance $J = (I, \text{sim})$ of some k -sound I'_0 , such that J satisfies Σ_{FD} and is $(B |I|)$ -envelope-saturated. Now, apply the Envelope-Thrifty

Completion Proposition to obtain an aligned superinstance J' of I_0 . By the Envelope Blowup Lemma, J' contains $< B|I|$ new facts, so, by Lemma VII.9, J' must still be 1-envelope-saturated. Hence, J' satisfies Σ . This concludes the proof, as J' is an aligned superinstance of I_0 .

VIII. Cyclic Queries

We now finally complete our proof of the Universal Models Theorem (Theorem III.5) by moving from acyclic Boolean CQs to arbitrary Boolean CQs. We do so by a generic process which is essentially independent from our previous construction.

Intuitively, the only cyclic CQs that hold in $\text{Chase}(I_0, \Sigma_{\text{UID}})$ either have an acyclic self-homomorphic match (so they are implied by an acyclic CQ that also holds) or have all cycles matched to elements of I_0 . Hence, in a k -sound instance for CQ, no other cyclic queries must be true. We ensure this by a cycle blowup process that takes the product of our I with a group of high girth, following Otto [14]. However, we need to adjust this construction to avoid creating FD violations.

We let $J_f = (I_f, \text{sim})$ be the aligned superinstance obtained from the Acyclic Universal Models Theorem (Theorem VII.1). Its underlying instance I_f is a finite superinstance of I_0 that satisfies Σ , and the k -bounded simulation sim guarantees that I_f is k -sound for ACQ. Our goal in this section is to make I_f k -sound for CQ while still satisfying Σ , so that it is k -universal. This will conclude the proof of the Universal Models Theorem (Theorem III.5).

VIII.1. Simple product

Let us first introduce preliminary notions:

Definition VIII.1. A group $G = (S, \cdot)$ over a finite set S consists of an associative **product law** $\cdot : S^2 \rightarrow S$, a **neutral element** $e \in S$, and an **inverse law** $\cdot^{-1} : S \rightarrow S$ such that $x \cdot x^{-1} = x^{-1} \cdot x = e$ for all $x \in S$. We say that G is **generated** by $X \subseteq S$ if all elements of S can be written as a product of elements of X and $X^{-1} := \{x^{-1} \mid x \in X\}$.

Given a group G generated by X , the **girth** of G under X is the length of the shortest non-empty word w of elements of X and X^{-1} such that $w_1 \cdots w_n = e$ and $w_i \neq w_{i+1}^{-1}$ for all $1 \leq i < n$. (If $X = \{g\}$ with $g = g^{-1}$, the girth is 1.)

Lemma VIII.2 ([12]). For all $n \in \mathbb{N}$ and finite non-empty set X , there is a finite group $G = (S, \cdot)$ generated by X with girth $\geq n$ under X . We call G an **n -acyclic group generated by X** .

In other words, in an n -acyclic group generated by X , there is no short product of elements of X and their inverses which evaluates to e , except those that include a factor xx^{-1} .

We now take the product of I_f with such a finite group G . This ensures that any cycles in the product instance are large, because they project to cycles in G . We use a specific generator:

Definition VIII.3. The **fact labels** of a superinstance I of I_0 are $\Lambda(I) := \{1_i^F \mid F \in I \setminus I_0, 1 \leq i \leq |F|\}$.

Now, we define the product of a superinstance I of I_0 with a group generated by $\Lambda(I)$. We make sure not to blow up cycles in I_0 , so the result remains a superinstance of I_0 :

Definition VIII.4. Let I be a finite superinstance of I_0 and G be a finite group generated by $\Lambda(I)$. The **product of I by G preserving I_0** is the finite instance $(I, I_0) \otimes G$ with domain $\text{dom}(I) \times G$ consisting of the following facts, for all $g \in G$:

- For every fact $R(\mathbf{a})$ of I_0 , the fact $R((a_1, g), \dots, (a_{|R|}, g))$.

- For every fact $F = R(\mathbf{a})$ of $I \setminus I_0$, the following fact:

$$R((a_1, g \cdot 1_1^F), \dots, (a_{|R|}, g \cdot 1_{|R|}^F)).$$

We identify (a, e) to a for $a \in \text{dom}(I_0)$, so $(I, I_0) \otimes G$ is still a superinstance of I_0 .

We say a superinstance I of I_0 is **k -instance-sound** (for Σ) if for any CQ q such that $|q| \leq k$, if q has a match in I involving an element of I_0 , then $\text{Chase}(I_0, \Sigma_{\text{UID}}) \models q$. We can ensure that I_f is k -instance-sound, up to having performed k chase rounds on I_0 initially. We can then state the following property:

Lemma VIII.5 (Simple product). *Let I be a finite superinstance of I_0 and G a finite $(2k+1)$ -acyclic group generated by $\Lambda(I)$. If I is k -sound for ACQ and k -instance-sound, then $(I, I_0) \otimes G$ is k -sound for CQ.*

Example VIII.6. Consider $F_0 := R(a, b)$, $I_0 := \{F_0\}$, and Σ_{UID} consisting of $\tau : R^2 \subseteq S^1$, $\tau' : S^2 \subseteq R^1$, τ^{-1} , and $(\tau')^{-1}$. Let $F := S(b, a)$, and $I := I_0 \sqcup \{F\}$. I satisfies Σ_{UID} and is sound for ACQ, but not for CQ: take for instance $q : \exists xy R(x, y) \wedge S(y, x)$, which is cyclic and holds in I while $(I_0, \Sigma_{\text{UID}}) \not\models_{\text{unr}} q$.

We have $\Lambda(I) = \{1_1^F, 1_2^F\}$. Identify 1_1^F and 1_2^F to 1 and 2 and consider the group $G := (\{0, 1, 2\}, \cdot)$ where \cdot is addition modulo 3. G has girth 2 under $\Lambda(I)$.

The product $I_p := (I, I_0) \otimes G$, writing pairs as subscripts for brevity, is $\{R(a_0, b_0), R(a_1, b_1), R(a_2, b_2), S(b_1, a_2), S(b_2, a_0), S(b_0, a_1)\}$. In this case I_p happens to be 5-sound for CQ.

We cannot conclude directly with the simple product, because $I_p := (I_f, I_0) \otimes G$ may violate Σ_{UFD} even though $I_f \models \Sigma_{\text{FD}}$. Indeed, there may be a relation R , a UFD $\phi : R^p \rightarrow R^q$ in Σ_{UFD} , and two R -facts F and F' in $I_f \setminus I_0$ with $\pi_{R^p, R^q}(F) = \pi_{R^p, R^q}(F')$. In I_p the images of F and F' may overlap only on R^p , so they could violate ϕ .

VIII.2. Mixed product

What we need is a more refined notion of product, that does not attempt to blow up cycles within fact overlaps. To define it, we need to consider a *quotient* of I_f :

Definition VIII.7. The *quotient* I/\sim of an instance I by an equivalence relation \sim on $\text{dom}(I)$ is defined as follows:

- $\text{dom}(I/\sim)$ is the equivalence classes of \sim on $\text{dom}(I)$,
- I/\sim contains one fact $R(\mathbf{A})$ for every fact $R(\mathbf{a})$ of I , where A_i is the \sim -class of a_i for all $R^i \in \text{Pos}(R)$.

The *quotient homomorphism* χ_\sim is the homomorphism from I to I/\sim defined accordingly.

We quotient I_f by the equivalence relation \simeq_k (recall Definition V.1), yielding $I'_f := I_f/\simeq_k$. The resulting I'_f may no longer satisfy Σ . However, it is still k -sound for ACQ, for the following reason:

Lemma VIII.8. Any k -bounded simulation from an instance I to an instance I' defines a k -bounded simulation from I/\simeq_k to I' .

We then consider the homomorphism χ_{\simeq_k} from I_f to I'_f , and blow up cycles in I_f by a *mixed product* that only distinguishes facts with a different image in I'_f by χ_{\simeq_k} . The point is that, as we show from our construction, facts of I_f that have the same elements at the same positions always have the same \simeq_k -class. Hence, they are mapped to the same fact by χ_{\simeq_k} and will not be distinguished by the mixed product. Let us formalize this:

Definition VIII.9. Let I be a superinstance of I_0 and h be a homomorphism from I to some instance I' . We say I is **cautious** for h (and I_0) if for any relation R , for any two R -facts F and F' such that $\pi_{R^p}(F) = \pi_{R^p}(F')$ for some $R^p \in \text{Pos}(R)$, either $F, F' \in I_0$, or $h(F) = h(F')$.

Lemma VIII.10 (Cautiousness). *The superinstance I_f of I_0 constructed by the Acyclic Universal Models Theorem (Theorem VII.1) is cautious for χ_{\simeq_k} .*

The reason why I_f is cautious for $h := \chi_{\simeq_k}$ is that, except for facts of I_0 , overlaps between facts only occur when reusing envelope elements at non-dangerous positions, in which case the sim-images of both facts are \simeq_k -equivalent in $\text{Chase}(I_0, \Sigma_{\text{UID}})$. We can then show that, from our construction, such elements are actually \simeq_k -equivalent in I_f .

We now define the notion of mixed product, which uses the same fact label for facts with the same image by h :

Definition VIII.11. Let I be a finite superinstance of I_0 with a homomorphism h to another finite superinstance I' of I_0 such that $h|_{I_0}$ is the identity and $h|_{(I \setminus I_0)}$ maps to $I' \setminus I_0$. Let G be a finite group generated by $\Lambda(I')$.

The **mixed product** of I by G via h preserving I_0 , written $(I, I_0) \otimes^h G$, is the finite superinstance of I_0 with domain $\text{dom}(I) \times G$ consisting of the following facts, for every $g \in G$:

- For every fact $R(\mathbf{a})$ of I_0 , the fact $R((a_1, g), \dots, (a_{|R|}, g))$.
- For every fact $R(\mathbf{a})$ of $I \setminus I_0$, the following fact:
 $R((a_1, g \cdot I_1^{h(F)}), \dots, (a_{|R|}, g \cdot I_{|R|}^{h(F)}))$.

We now show that the mixed product preserves UIDs and FDs when cautiousness is assumed.

Lemma VIII.12 (Mixed product preservation). *For any UID or FD τ , if $I \models \tau$ and I is cautious for h , then $(I, I_0) \otimes^h G \models \tau$.*

Second, we show that $h : I \rightarrow I'$ lifts to a homomorphism from the mixed product to the simple product.

Lemma VIII.13 (Mixed product homomorphism). *There is a homomorphism from $(I, I_0) \otimes^h G$ to $(I', I_0) \otimes G$ which is the identity on $I_0 \times G$.*

We can now conclude our proof of the Universal Models Theorem (Theorem III.5). We construct $J_f = (I_f, \text{sim})$ by the Acyclic Universal Models Theorem (Theorem VII.1) and consider I_f . It is a finite superinstance of I_0 which is k -universal for Σ and ACQ. Further, up to having distinguished the elements of I_0 with fresh predicates and having performed initial chasing, we can ensure that $I'_f := I_f / \simeq_k$ is k -instance-sound and that the homomorphism $\chi_{\simeq_k} : I_f \rightarrow I'_f$ satisfies the hypotheses of the mixed product.

Let G be a $(2k + 1)$ -acyclic group generated by $\Lambda(I'_f)$, and consider $I_p := (I'_f, I_0) \otimes G$. As I_f was k -sound for ACQ, so is I'_f by Lemma VIII.8, and as I'_f is also k -instance-sound, I_p is k -sound for CQ by the Simple Product Lemma (Lemma VIII.5). However, as we explained, in general $I_p \not\models \Sigma$. We thus construct $I_m := (I_f, I_0) \otimes^h G$, with $h := \chi_{\simeq_k}$. By the Mixed Product Homomorphism Lemma, I_m has a homomorphism to I_p , so it is also k -sound for CQ. Further, I_f is cautious for χ_{\simeq_k} by the Cautiousness Lemma, so, by the Mixed Product Preservation Lemma, we have $I_m \models \Sigma$ because $I_f \models \Sigma$.

Hence, the mixed product I_m is a finite k -universal instance for Σ and CQ. This concludes the proof of the Universal Models Theorem, and hence of our main theorem (Theorem III.2).

IX. Conclusion

In this work we have developed the first techniques on arbitrary arity schemas to build finite models that satisfy both referential constraints and number restrictions, while controlling which CQs are satisfied. We have used this to prove that finite open-world query answering for CQs, UIDs and FDs is finitely controllable up to finite closure of the dependencies. Using this, we have isolated the complexity of FQA for UIDs and FDs.

As presented the constructions are quite specific to dependencies, but in future work we will look to extend them to constraint languages containing disjunction, with the goal of generalizing to higher arity the rich arity-2 constraint languages of, e.g., [10, 15], while maintaining the decidability of FQA.

Acknowledgements. This work was supported in part by the Engineering and Physical Sciences Research Council, UK (EP/G004021/1) and the French ANR NormAtis project. We are very grateful to Balder ten Cate, Thomas Gogacz, Andreas Pieris, and Pierre Senellart for comments on earlier drafts, and to the anonymous reviewers of LICS for their valuable feedback.

References

- [1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.
- [2] W. W. Armstrong. Dependency structure of data base relationships. In *IFIP Congress*, 1974.
- [3] V. Bárány, G. Gottlob, and M. Otto. Querying the guarded fragment. In *LICS*, 2010.
- [4] A. Cali, G. Gottlob, and A. Pieris. Towards more expressive ontology languages: The query answering problem. *Artif. Intel.*, 193, 2012.
- [5] A. Cali, D. Lembo, and R. Rosati. On the decidability and complexity of query answering over inconsistent and incomplete databases. In *PODS*, 2003.
- [6] A. Cali, D. Lembo, and R. Rosati. Query rewriting and answering under constraints in data integration systems. In *IJCAI*, 2003.
- [7] M. A. Casanova, R. Fagin, and C. H. Papadimitriou. Inclusion dependencies and their interaction with functional dependencies. *JCSS*, 28(1), 1984.
- [8] S. S. Cosmadakis, P. C. Kanellakis, and M. Y. Vardi. Polynomial-time implication problems for unary inclusion dependencies. *JACM*, 37(1), 1990.
- [9] R. Fagin, P. G. Kolaitis, R. J. Miller, and L. Popa. Data exchange: Semantics and query answering. In *ICDT*, 2003.
- [10] Y. Ibáñez-García, C. Lutz, and T. Schneider. Finite model reasoning in Horn description logics. In *KR*, 2014.
- [11] D. S. Johnson and A. C. Klug. Testing containment of conjunctive queries under functional and inclusion dependencies. *JCSS*, 28(1), 1984.
- [12] G. A. Margulis. Explicit constructions of graphs without short cycles and low density codes. *Combinatorica*, 2:71–78, 1982.

- [13] A. Onet. The chase procedure and its applications in data exchange. In *Data Exchange, Information, and Streams*, 2013.
- [14] M. Otto. Modal and guarded characterisation theorems over finite transition systems. In *LICS*, 2002.
- [15] I. Pratt-Hartmann. Data-complexity of the two-variable fragment with counting quantifiers. *Inf. Comput.*, 207(8), 2009.
- [16] R. Rosati. On the decidability and finite controllability of query processing in databases with incomplete information. In *PODS*, 2006.
- [17] R. Rosati. Finite model reasoning in DL-Lite. In *ESWC*, 2008.
- [18] R. Rosati. On the finite controllability of conjunctive query answering in databases under open-world assumption. *JCSS*, 77(3), 2011.

A. Details about the UID chase and Unique Witness Property

Recall the *Unique Witness Property*:

For any element $a \in \text{dom}(\text{Chase}(I, \Sigma_{\text{UID}}))$ and position R^p of σ , if two facts of $\text{Chase}(I, \Sigma_{\text{UID}})$ contain a at position R^p , then they are both facts of I .

We first exemplify why this may not be guaranteed by the first round of the UID chase. Consider the instance $I = \{R(a), S(a)\}$ and the UIDs $\tau_1 : R^1 \subseteq T^1$ and $\tau_2 : S^1 \subseteq T^1$, where T is binary. Applying a round of the UID chase creates the instance $\{R(a), S(a), T(a, b_1), T(a, b_2)\}$, with $T(a, b_1)$ being created by applying τ_1 to the active fact $R(a)$, and $T(a, b_2)$ being created by applying τ_2 to the active fact $S(a)$.

By contrast, the core chase would create only one of these two facts, because it would consider that two new facts are *equivalent*: they have the same exported element occurring at the same position. In general, the core chase keeps only one fact within each class of equivalent facts.

However, after one chase round by the core chase, there is no longer any distinction between the UID chase and the core chase, because the following property holds on the result I' of a chase round (by the core chase or the UID chase) on any instance I'' : (*) for any $\tau \in \Sigma_{\text{UID}}$ and element $a \in \text{Wants}(I', \tau)$, a occurs in only one fact of I' . This is true because Σ_{UID} is transitively closed, so we know that no UID of Σ_{UID} is applicable to an element of $\text{dom}(I'')$ in I' ; hence the only elements that witness violations occur in the one fact where they were introduced in I' .

We now claim that (*) implies the Unique Witness Property. Indeed, assume to the contrary that $a \in \text{dom}(\text{Chase}(I, \Sigma_{\text{UID}}))$ violates it.

If $a \in \text{dom}(I)$, because Σ_{UID} is transitively closed, after the first chase round on I , we no longer create any fact that involves a . Hence, each one of F_1 and F_2 is either a fact of I or a fact created in the first round of the chase (which is a chase round by the core chase). However, if one of F_1 and F_2 is in I , then it witnesses that we could not have $a \in \text{Wants}(I, R^p)$, so it is not possible that the other fact was created in the first chase round. It cannot be the case either that F_1 and F_2 were both created in the first chase round, by definition of the core chase. Hence, F_1 and F_2 are necessarily both facts of I .

If $a \in \text{dom}(\text{Chase}(I, \Sigma_{\text{UID}})) \setminus \text{dom}(I)$, assume that a occurs at position R^p in two facts F_1, F_2 . As $a \notin \text{dom}(I)$, none of them is a fact of I . We then show a contradiction. It is not possible that one of those facts was created in a chase round before the other, as otherwise the second created fact could not have been created because of the first created fact. Hence, both facts must have been created in

the same chase round. So there was a chase round from I'' to I' where we had $a \in \text{Wants}(I'', R^p)$ and both F_1 and F_2 were created respectively from active facts F'_1 and F'_2 of I'' by UIDs $\tau_1 : S^q \subseteq R^p$ and $\tau_2 : T^r \subseteq R^p$. But then, by property (*), a occurs in only one fact, so as it occurs in F'_1 and F'_2 we have $F'_1 = F'_2$. Further, as $a \notin \text{dom}(I)$, F'_1 and F'_2 are not facts of I either, so by definition of the UID chase and of the core chase, it is easy to see a occurs at only one position in $F'_1 = F'_2$. This implies that $\tau_1 = \tau_2$. Hence, we must have $F_1 = F_2$.

B. Proofs for Section III: Main Result and Overall Approach

B.1. Proof of Proposition III.1 (Complexity of UQA for FDs and UIDs)

Proposition III.1. *UQA for FDs and UIDs has PTIME data complexity and NP-complete combined complexity.*

We first show the results for UIDs in isolation. UQA for UIDs is NP-complete in combined complexity: the lower bound is immediate from query evaluation [1], the upper bound is by Johnson & Klug [11] and actually holds for IDs of arbitrary fixed arity (which they call “width”). For data complexity, Calì et al. [6] showed a PTIME (in fact, AC^0) upper bound for arbitrary IDs by observing that the certain answers can be expressed by another first-order query.

We now show that the same upper bounds apply to UQA for UIDs and FDs (the lower bound clearly also applies). This result is implicit in prior work of [5, 4], but we prove it here for completeness. We argue that UIDs and FDs are *separable*. This means that for any conjunction Σ of FDs Σ_{FD} and UIDs Σ_{UID} , for any instance I_0 and CQ q , if $I_0 \models \Sigma_{\text{FD}}$ then we have $(I_0, \Sigma) \models_{\text{unr}} q \leftrightarrow (I_0, \Sigma_{\text{UID}}) \models_{\text{unr}} q$. From this result, the upper bounds follow from the bounds for the UID case above, since checking whether $I_0 \models \Sigma_{\text{FD}}$ can be done in PTIME. Separability follows from the *non-conflicting condition* of [5, 4] but we give a simpler argument.

Assume that I_0 satisfies Σ_{FD} . Clearly if $(I_0, \Sigma_{\text{UID}}) \models_{\text{unr}} q$ then $(I_0, \Sigma) \models_{\text{unr}} q$. We thus need to show that if $(I_0, \Sigma) \models_{\text{unr}} q$ then $(I_0, \Sigma_{\text{UID}}) \models_{\text{unr}} q$. Consider $\text{Chase}(I_0, \Sigma_{\text{UID}})$. If $\text{Chase}(I_0, \Sigma_{\text{UID}}) \models \Sigma_{\text{FD}}$, then $\text{Chase}(I_0, \Sigma_{\text{UID}})$ is a superinstance of I_0 that satisfies Σ , so because $(I_0, \Sigma) \models_{\text{unr}} q$ we must have $\text{Chase}(I_0, \Sigma_{\text{UID}}) \models q$. By universality of the chase, this implies $(I_0, \Sigma_{\text{UID}}) \models_{\text{unr}} q$.

Hence, it suffices to show that $\text{Chase}(I_0, \Sigma_{\text{UID}}) \models \Sigma_{\text{FD}}$. Assume to the contrary the existence of F and F' in $\text{Chase}(I_0, \Sigma_{\text{UID}})$ violating an FD of Σ_{FD} . There must exist a position $R^p \in \text{Pos}(\sigma)$ such that $\pi_{R^p}(F) = \pi_{R^p}(F')$. By the Unique Witness Property, this implies that F and F' are facts of I_0 , which is impossible by our assumption that $I_0 \models \Sigma_{\text{FD}}$.

B.2. Proof of the Main Theorem (Theorem III.2) from the Universal Models Theorem (Theorem III.5)

To show the Main Theorem from the Universal Models Theorem, let Σ be a conjunction of FDs and UIDs, Σ' its finite closure, and I_0 a finite instance. We want to show finite controllability up to finite closure, namely, $(I_0, \Sigma) \models_{\text{fin}} q$ iff $(I_0, \Sigma') \models_{\text{unr}} q$.

We can assume without loss of generality that I_0 satisfies the FDs of Σ' , as otherwise there is no superinstance of I_0 satisfying Σ' , and both problems are always vacuously true.

It is clear that for any CQ q , we have $(I_0, \Sigma) \models_{\text{fin}} q$ iff $(I_0, \Sigma') \models_{\text{fin}} q$. Indeed, Σ' includes Σ and conversely any finite superinstance of I_0 which satisfies Σ must satisfy Σ' , by definition of the finite closure. So in fact, to prove finite controllability up to finite closure, it suffices to show that $(I_0, \Sigma') \models_{\text{fin}} q$

q iff $(I_0, \Sigma') \models_{\text{unr}} q$ for any CQ q . The backward implication is immediate as all finite superinstances of I_0 satisfying Σ' are also unrestricted superinstances. We prove the contrapositive of the forward implication.

Let q be a CQ, let $k := |q|$, and assume that $(I_0, \Sigma') \not\models_{\text{unr}} q$. By the Universal Models Theorem, let I be a finite superinstance of I_0 that is $|q|$ -sound and satisfies Σ' . As I is $|q|$ -sound, we have $I \not\models q$, so, as I is a finite superinstance of I_0 that satisfies Σ' , it witnesses that $(I_0, \Sigma') \not\models_{\text{fin}} q$. This proves the desired equivalence. Hence, we have established that Σ' is finitely controllable up to finite closure, and have proved the Main Theorem.

B.3. Proof of Corollary III.3 (Complexity of FQA for FDs and UIDs)

Corollary III.3. *FQA for FDs and UIDs has PTIME data complexity and NP-complete combined complexity.*

By our Main Theorem (Theorem III.2), any instance (I, Σ, q) to the FQA problem, formed of an instance I , a conjunction Σ of IDs Σ_{UID} and FDs Σ_{FD} , and a CQ q , reduces to the UQA instance (I, Σ', q) , where Σ' is the finite closure of Σ . Computing Σ' from Σ is data-independent, so the PTIME data complexity result of Proposition III.1 clearly still applies. It is also clear that the NP-hardness combined complexity bound of Proposition III.1 can be re-proven for FQA, as it already held even when $\Sigma = \emptyset$. So we only need to show that the combined complexity of FQA is in NP. A naive approach would be to compute explicitly Σ' and solve the UQA instance I, Σ', q ; but materializing Σ' may take exponential time.

Instead, remember that from our study of UQA complexity in the proof of Proposition III.1, UQA for UIDs and FDs can be performed by first checking the FDs on the initial instance, and then performing UQA for the UIDs in isolation. Hence, let Σ'_{UID} and Σ'_{FD} be the UIDs and FDs of Σ' . Rather than materializing Σ' , we will show that we can decide whether $I \models \Sigma'_{\text{FD}}$ in PTIME, and compute Σ'_{UID} in PTIME, which suffices to prove the claim as the combined complexity of deciding whether $(I, \Sigma'_{\text{UID}}) \models_{\text{unr}} q$ is then in NP.

We first justify that we can indeed compute Σ'_{UID} in PTIME. We consider every possible UID on positions occurring in Σ (there are polynomially many), and for each of them, determine in PTIME from Σ whether it is in Σ' , using the implication procedure of Cosmadakis et al. [8]. This allows us to compute Σ'_{UID} in PTIME.

We next justify that we can decide whether $I \models \Sigma'_{\text{FD}}$ in PTIME. For the same reason as for the UIDs, we can compute in PTIME from Σ the set Σ'_{UFD} of the UFDs which are in Σ' , by deciding implication for each possible UFD. We now argue that to test whether $I \models \Sigma'_{\text{FD}}$, it suffices to test whether $I \models \Sigma_{\text{FD}}$ and whether $I \models \Sigma'_{\text{UFD}}$. This follows if we can show that Σ'_{FD} is implied by $\Sigma'_{\text{UFD}} \cup \Sigma_{\text{FD}}$ by the usual axiomatization of unrestricted and finite implication for FDs alone, from Armstrong [2]. Indeed, in this case, if $I \models \Sigma'_{\text{FD}}$ then $I \models \Sigma'_{\text{UFD}} \cup \Sigma_{\text{FD}}$ as it is a subset of Σ'_{FD} , and conversely if $I \models \Sigma'_{\text{UFD}} \cup \Sigma_{\text{FD}}$ then I satisfies Σ'_{FD} because they are implied by $\Sigma'_{\text{UFD}} \cup \Sigma_{\text{FD}}$ so are also satisfied by any instance that satisfies $\Sigma'_{\text{UFD}} \cup \Sigma_{\text{FD}}$.

To justify that Σ'_{FD} is implied by $\Sigma'_{\text{UFD}} \cup \Sigma_{\text{FD}}$, we use Theorem 4.1 of [8], according to which a sound and complete axiomatization of the finite closure of FDs and UIDs consists of the usual FD implication rules, the standard UID axiomatization of Casanova et al. [7], and the *cycle rule*. So, consider any FD ϕ of Σ'_{FD} and let us justify that it is implied by $\Sigma'_{\text{UFD}} \cup \Sigma_{\text{FD}}$. If ϕ is a UFD, then $\phi \in \Sigma'_{\text{UFD}}$. Otherwise the last steps of a derivation of ϕ with the axiomatization of [8] must be rules from the FD implication rules, as they are the only ones which can deduce higher-arity FDs. Let us group together the last FD implication rules that were applied, and consider the set S of the hypotheses to FD implication rules

that were not themselves produced by FD implication rules. Each hypothesis from S is either an FD of Σ_{FD} or was produced by the cycle rule. Now, the cycle rule can only deduce UFDs (and UIDs). Hence, $S \subseteq \Sigma_{\text{FD}} \cup \Sigma'_{\text{UFD}}$, which implies that we can construct a derivation of ϕ from $\Sigma_{\text{FD}} \cup \Sigma'_{\text{UFD}}$ using the FD implication rules. Thus, we can indeed compute in PTIME $\Sigma'_{\text{UFD}} \cup \Sigma_{\text{FD}}$, and check in PTIME whether $I \models \Sigma'_{\text{UFD}} \cup \Sigma_{\text{FD}}$, and we have shown that this is equivalent to checking whether $I \models \Sigma'_{\text{FD}}$. This concludes the proof.

C. Proofs for Section IV: Weak-Soundness and Reversible UIDs

This section proves the Acyclic Unary Weakly-Sound Models Proposition (Proposition IV.2), which weakens the Acyclic Unary Models Theorem (Theorem III.6) by making assumption reversible and replacing k -soundness by weak-soundness (Definition IV.1).

C.1. Proof of Proposition IV.4 (Satisfying UIDs in balanced instances)

Proposition IV.4. *Assuming binary and reversible, any balanced finite instance I satisfying Σ_{UFD} has a finite weakly-sound superinstance I' that satisfies Σ_{U} , with $\text{dom}(I') = \text{dom}(I)$.*

For every relation R of σ , let f_R be a bijection between $\text{Wants}(I, R^1)$ and $\text{Wants}(I, R^2)$; this is possible, because I is balanced.

Consider the superinstance I' of I , with $\text{dom}(I') = \text{dom}(I)$, obtained by adding, for every R of σ , the fact $R(a, f_R(a))$ for every $a \in \text{Wants}(I, R^1)$. I' is clearly a finite weakly-sound superinstance of I , because for every $a \in \text{dom}(I')$, if a occurs at some position R^p in some fact F of I' , then either F is a fact of I and $a \in \pi_{R^p}(I)$, or F is a new fact and by definition $a \in \text{Wants}(I, R^p)$.

Let us show that $I' \models \Sigma_{\text{UFD}}$. Assume to the contrary that there are two facts F and F' in I' that witness a violation of a UFD $\phi : R^p \rightarrow R^q$ of Σ_{UFD} . As $I \models \Sigma_{\text{UFD}}$, one of F and F' is necessarily a new fact; we assume without loss of generality that it is F . Consider $a := \pi_{R^p}(F)$. By definition of the new facts, we have $a \in \text{Wants}(I, R^p)$, so that $a \notin \pi_{R^p}(I)$. Now, as $\{F, F'\}$ is a violation, we must have $\pi_{R^p}(F) = \pi_{R^p}(F')$, so as $a \notin \pi_{R^p}(I)$, F' must also be a new fact. Hence, by definition of the new facts, letting $b := \pi_{R^q}(F)$ and $b' := \pi_{R^q}(F')$, depending on whether $p = 1$ or $p = 2$ we have either $b = b' = f_R(a)$ or $b = b' = f_R^{-1}(a)$, which is well-defined because f_R is a bijection. This contradicts the fact that F and F' violate ϕ .

Let us now show that $I' \models \Sigma_{\text{UID}}$. Assume to the contrary that there is an active fact $F = R(a_1, a_2)$, for a UID $\tau : R^p \subseteq S^q$. If F is a fact of I , we had $a_p \in \text{Wants}(I, S^q)$, so F cannot be an active fact in I' by construction of f_S . So we must have $F \in I' \setminus I$. Hence, by definition of the new facts, we had $a_p \in \text{Wants}(I, R^p)$; so there must be $\tau' : T^r \subseteq R^p$ in Σ_{UID} such that $a_p \in \pi_{T^r}(I)$. Hence, because Σ_{UID} is transitively closed, either $T^r = S^q$ or the UID $T^r \subseteq S^q$ is in Σ_{UID} . In the first case, as $a_p \in \pi_{T^r}(I)$, F cannot be an active fact for τ , a contradiction. In the second case, we had $a_p \in \text{Wants}(I, S^q)$, which is a contradiction for the same reason as before.

Hence, I' is a finite weakly-sound superinstance of I that satisfies Σ_{U} and with $\text{dom}(I') = \text{dom}(I)$, the desired claim.

C.2. Proof of the Balancing Lemma (Lemma IV.9)

Lemma IV.9 (Balancing). *For any finite instance I , if I satisfies Σ_{UFD} then it has a balanced pssinstance.*

We prove the lemma without assumption binary, as we will use it without this assumption later in Section IV.

For any position R^p define $o(R^p) := \text{Wants}(I, R^p) \sqcup \pi_{R^p}(I)$. Intuitively, those are the elements that either appear at R^p or want to appear there. We claim that $o(R^p) = o(S^q)$ whenever $R^p \sim_{\text{ID}} S^q$. Indeed, we have $\pi_{R^p}(I) \subseteq o(S^q)$: elements in $\pi_{R^p}(I)$ want to appear at S^q unless they already do, and in both cases they are in $o(S^q)$. Likewise, elements of $\text{Wants}(I, R^p)$ either occur at S^q , or at some other position T^r such that $T^r \subseteq R^p$ is a UID of Σ_{UID} , so that by transitivity $T^r \subseteq S^q$ also is, and so they want to be at S^q unless they already are. Hence $o(R^p) \subseteq o(S^q)$, and symmetrically $o(S^q) \subseteq o(R^p)$.

Let $N := \max_{R^p \in \text{Pos}(\sigma)} |o(R^p)|$, which is finite. We write $[R^p]_{\text{ID}}$ the \sim_{ID} -class of any position R^p . We define for each \sim_{ID} -class $[R^p]_{\text{ID}}$ a set $p([R^p]_{\text{ID}})$ of $N - |o(R^p)|$ fresh values. We let \mathcal{H} be the disjoint union of the $p([R^p]_{\text{ID}})$ for all classes $[R^p]_{\text{ID}}$, and set λ to map the elements of $p([R^p]_{\text{ID}})$ to $[R^p]_{\text{ID}}$. We have thus defined our pssinstance $P = (I, \mathcal{H}, \lambda)$.

Let us now show that P is balanced. Consider now two positions R^p and R^q such that $\phi : R^p \rightarrow R^q$ and $\phi' : R^q \rightarrow R^p$ are in Σ_{UFD} , and show that $|\text{Wants}(P, R^p)| = |\text{Wants}(P, R^q)|$. We have $|\text{Wants}(P, R^p)| = |\text{Wants}(I, R^p)| + |p([R^p]_{\text{ID}})| = |o(R^p)| - |\pi_{R^p}(I)| + N - |o(R^p)|$, which simplifies to $N - |\pi_{R^p}(I)|$. Similarly $|\text{Wants}(P, R^q)| = N - |\pi_{R^q}(I)|$. Since $I \models \Sigma_{\text{UFD}}$ and ϕ and ϕ' are in Σ_{UFD} we know that $|\pi_{R^p}(I)| = |\pi_{R^q}(I)|$. From this the conclusion follows.

C.3. Proof of the Binary Realizations Lemma (Lemma IV.10)

Lemma IV.10 (Binary realizations). *For any balanced pssinstance P of an instance I that satisfies Σ_{UFD} , we can construct a realization of P that satisfies Σ_{U} .*

Let us construct a realization I' of P . We construct bijections f_R for every relation R between $\text{Wants}(P, R^1)$ and $\text{Wants}(P, R^2)$ as for Proposition IV.4; this is possible, as P is balanced. We then construct I' in the same way, by adding to I , for every R of σ , the fact $R(a, f_R(a))$ for every $a \in \text{Wants}(P, R^1)$.

We prove that I' is a realization again by observing that whenever we create a fact $R(a, f_R(a))$, then we have $a \in \text{Wants}(P, R^1)$ and $f_R(a) \in \text{Wants}(P, R^2)$.

The fact that I' satisfies Σ_{UFD} is for the same reason as for Proposition IV.4.

We now show that I' satisfies Σ_{UID} . Assume to the contrary that there is an active fact $F = R(a_1, a_2)$, for a UID $\tau : R^p \subseteq S^q$, so that $a_p \in \text{Wants}(I', R^p)$. If $a_p \in \text{dom}(I)$, then the proof is exactly as for Proposition IV.4. Otherwise, if $a_p \in \mathcal{H}$, clearly by construction of f_R and I' we have $a_p \in \pi_{T^r}(I')$ iff $T^r \in \lambda(a_p)$. Hence, as $a_p \in \pi_{R^p}(I')$ and as τ witnesses by assumption reversible that $R^p \sim_{\text{ID}} S^q$, we have $a_p \in \pi_{S^q}(I')$, contradicting the fact that $a_p \in \text{Wants}(I', S^q)$.

C.4. Proof of Lemma “Binary realizations are completions” (Lemma IV.11)

Lemma IV.11 (Binary realizations are completions). *If I' is a realization of a pssinstance of I then it is a weakly-sound superinstance of I .*

Clearly I' is a superinstance of I . Let us show that it is weakly-sound. Recall the definition of a weakly-sound superinstance:

Definition IV.1. A superinstance I' of an instance I is **weakly-sound** if the following holds:

- for any $a \in \text{dom}(I)$ and $R^p \in \text{Pos}(\sigma)$, if $a \in \pi_{R^p}(I')$, then either $a \in \pi_{R^p}(I)$ or $a \in \text{Wants}(I, R^p)$;
- for any $a \in \text{dom}(I') \setminus \text{dom}(I)$ and $R^p, S^q \in \text{Pos}(\sigma)$, if $a \in \pi_{R^p}(I')$ and $a \in \pi_{S^q}(I')$ then $R^p = S^q$ or $R^p \subseteq S^q$ is in Σ_{UID} .

Consider $a \in \text{dom}(I')$ and $R^p \in \text{Pos}(\sigma)$ such that $a \in \pi_{R^p}(I')$. As I' is a realization, we know that either $a \in \pi_{R^p}(I)$ or $a \in \text{Wants}(P, R^p)$. By definition of $\text{Wants}(P, R^p)$, and because $\mathcal{H} = \text{dom}(I') \setminus \text{dom}(I)$, this means that either $a \in \text{dom}(I)$ and $a \in \pi_{R^p}(I) \sqcup \text{Wants}(I, R^p)$, or $a \in \text{dom}(I') \setminus \text{dom}(I)$ and $R^p \in \lambda(a)$. Hence:

- For any $a \in \text{dom}(I)$ and $R^p \in \text{Pos}(\sigma)$, we have established that $a \in \pi_{R^p}(I')$ implied that either $a \in \pi_{R^p}(I)$ or $a \in \text{Wants}(I, R^p)$.
- For any $a \in \text{dom}(I') \setminus \text{dom}(I)$ and for any $R^p, S^q \in \text{Pos}(\sigma)$, we know that $R^p, S^q \in \lambda(a)$, which implies that $R^p \sim_{\text{ID}} S^q$, so $R^p = S^q$ or $R^p \subseteq S^q$ is in Σ_{UID} .

So indeed the two conditions of weak-soundness hold.

C.5. Proof of the Realizations Lemma (Lemma IV.16)

Lemma IV.16 (Realizations). *For any balanced pssinstance P of an instance I that satisfies Σ_{UFD} , we can construct a Σ_{U} -compliant piecewise realization of P .*

Let $P = (I, \mathcal{H}, \lambda)$ be the balanced pssinstance. Recall that the $\leftrightarrow_{\text{FUN}}$ -classes of σ are numbered Π_1, \dots, Π_n . By definition of being balanced (Definition IV.3), for any $\leftrightarrow_{\text{FUN}}$ -class Π_i , for any two positions $R^p, R^q \in \Pi_i$, we have $|\text{Wants}(P, R^p)| = |\text{Wants}(P, R^q)|$. Hence, for all $1 \leq i \leq n$, let s_i be the value of $|\text{Wants}(P, R^p)|$ for any $R^p \in \Pi_i$. For $1 \leq i \leq n$, we let m_i be the arity of Π_i , and number the positions of Π_i as $R^{p_1^i}, \dots, R^{p_{m_i}^i}$. We define for each $1 \leq i \leq n$ and $1 \leq j \leq m_i$ a bijection ϕ_j^i from $\{1, \dots, s_i\}$ to $\text{Wants}(P, R^{p_j^i})$. We construct the piecewise realization $PI = (K_1, \dots, K_n)$ by setting each K_i for $1 \leq i \leq n$ to be $\pi_{\Pi_i}(I)$ plus the tuples $(\phi_1^i(I), \dots, \phi_{m_i}^i(I))$ for $1 \leq l \leq s_i$.

It is clear that PI is indeed a piecewise realization, because whenever we create a tuple $\mathbf{a} \in \Pi_i$ for any $1 \leq i \leq n$, then, for any $R^p \in \Pi_i$, we have $a_p \in \text{Wants}(P, R^p)$.

Let us then show that PI is Σ_{UFD} -compliant. Assume by contradiction that there is $1 \leq i \leq n$ and $\mathbf{a}, \mathbf{b} \in K_i$ such that $a_l = b_l$ but $a_r \neq b_r$ for some $R^l, R^r \in \Pi_i$. As I satisfies Σ_{UFD} , we assume without loss of generality that $\mathbf{a} \in K_i \setminus \pi_{\Pi_i}(I)$. Now either $\mathbf{b} \in \pi_{\Pi_i}(I)$ or $\mathbf{b} \in K_i \setminus \pi_{\Pi_i}(I)$.

If $\mathbf{b} \in \pi_{\Pi_i}(I)$, then we know that $b_l \in \pi_{R^l}(I)$, but we know by construction that, as $\mathbf{a} \in K_i \setminus \pi_{\Pi_i}(I)$, we have $a_l \in \text{Wants}(P, R^l)$. Now, as $a_l = b_l$ and $b_l \in \text{dom}(I)$, we have $a_l \in \text{dom}(I)$, so that by definition of $\text{Wants}(P, R^l)$ we have $a_l \in \text{Wants}(I, R^l)$. Thus, as $a_l = b_l$, we have a contradiction.

Now, if $\mathbf{b} \in K_i \setminus \pi_{\Pi_i}(I)$, then, writing $R^l = R^{p_j^i}$ and $R^r = R^{p_{j'}^i}$, the fact that $a_l = b_l$ but $a_r \neq b_r$ contradicts the fact that $\phi_j^i \circ (\phi_{j'}^i)^{-1}$ is injective. Hence, PI is Σ_{UFD} -compliant.

Let us now show that PI is Σ_{UID} -compliant.

We must show that, for every UID $\tau : R^p \subseteq S^q$ of Σ_{UID} , we have $\text{Wants}(PI, \tau) = \emptyset$, which means that we have $\pi_{R^p}(PI) \subseteq \pi_{S^q}(PI)$. Let Π_i be the $\leftrightarrow_{\text{FUN}}$ -class of R^p , and assume to the contrary the existence of a tuple \mathbf{a} of K_i such that $a_p \notin \pi_{S^q}(PI)$. Either we have $a_p \in \text{dom}(I)$, or we have $a_p \in \mathcal{H}$.

In the first case, as $a_p \notin \pi_{S^q}(PI)$, in particular $a_p \notin \pi_{S^q}(I)$, and as $a_p \in \pi_{R^p}(I)$, we have $a_p \in \text{Wants}(I, \tau)$, so $a_p \in \text{Wants}(I, S^q)$. By construction of PI , then, letting i' be the $\leftrightarrow_{\text{FUN}}$ -class of S^q and letting $S^q = S^{p_{j'}^{i'}}$, as $\phi_{j'}^{i'}$ is surjective, we must have $a_p \in \pi_{S^q}(K_{i'})$, that is, $a_p \in \pi_{S^q}(PI)$, a contradiction.

In the second case, clearly by construction we have $a_p \in \pi_{T^r}(PI)$ iff $T^r \in \lambda(a_p)$, so that, given that τ witnesses $R^p \sim_{\text{ID}} S^q$, if $a_p \in \pi_{R^p}(PI)$ then $a_p \in \pi_{S^q}(PI)$, a contradiction.

We deduce that PI is indeed a Σ_{U} -compliant piecewise realization of P , completing the proof.

C.6. Proof of the Relation-Saturated Solutions Lemma (Lemma IV.20)

Lemma IV.20 (Relation-saturated solutions). *The result of performing sufficiently many chase rounds on any instance I is relation-saturated.*

Recall the definition of an instance being *relation-saturated*:

Definition IV.18. A relation R is **achieved** (by I and Σ_{UID}) if there is some R -fact in $\text{Chase}(I, \Sigma_{\text{UID}})$.

A superinstance I' of an instance I is **relation-saturated** (for Σ_{UID}) if every achieved relation (by I and Σ_{UID}) occurs in I' .

We now prove the lemma. For every relation R , either R is not achieved by I and Σ_{UID} , or there is $n_R \in \mathbb{N}$ such that there is a R -fact of $\text{Chase}(I, \Sigma_{\text{UID}})$ generated at the n_R -th round of the chase. Let $n := \max_{R \in \sigma} n_R$. As the number of relations in σ is finite, n is finite. Hence, letting I' be the result of applying n chase rounds to I , it is clear that I' is relation-saturated.

C.7. Proof of Lemma “Using realizations to get completions” (Lemma IV.21)

Lemma IV.21 (Using realizations to get completions). *For any finite relation-saturated instance I that satisfies Σ_{UFD} , from a Σ_{U} -compliant piecewise realization PI of a pssinstance of I , we can construct a finite weakly-sound superinstance of I that satisfies Σ_{U} .*

Recall that we number Π_1, \dots, Π_n the $\leftrightarrow_{\text{FUN}}$ -classes of $\text{Pos}(\sigma)$. We first define the following notion:

Definition C.1. We say that Π_j is an **inner** $\leftrightarrow_{\text{FUN}}$ -class if it contains a position occurring in Σ_{UID} ; otherwise, it is an **outer** $\leftrightarrow_{\text{FUN}}$ -class.

Intuitively, “outer” $\leftrightarrow_{\text{FUN}}$ -classes are those to which no UID of Σ_{UID} can apply, so we can create fresh elements at the positions of these classes without fear that UIDs will be applicable to the fresh elements.

We will use the notion of dangerous and non-dangerous positions from Section V:

Definition V.7. We say a position $S^r \in \text{Pos}(\sigma)$ is **dangerous** for a position $S^q \neq S^r$ if $S^r \rightarrow S^q$ is in Σ_{UFD} , and write $S^r \in \text{Dng}(S^q)$. Otherwise, S^r is **non-dangerous**, written $S^r \in \text{NDng}(S^q)$. Note that $\{S^q\} \sqcup \text{Dng}(S^q) \sqcup \text{NDng}(S^q) = \text{Pos}(S)$.

Observe that, if $R^p \leftrightarrow_{\text{FUN}} R^q$, then for $R^r \notin \{R^p, R^q\}$, we have $R^r \in \text{Dng}(R^p)$ iff $R^r \in \text{Dng}(R^q)$, and likewise for $\text{NDng}(R^p)$ and $\text{NDng}(R^q)$. So it makes sense to define $\text{Dng}(\Pi_i)$ or $\text{NDng}(\Pi_i)$, for Π_i an $\leftrightarrow_{\text{FUN}}$ -class of positions of some relation R , to refer to the positions of $\text{Pos}(R) \setminus \Pi_i$ that are dangerous or non-dangerous for some $R^p \in \Pi_i$ (and hence for all of them).

We show a first lemma about the positions where FD violations may be introduced:

Lemma C.2. For any relation R and FDs Σ_{FD} , for any $R^p \in \text{Pos}(R)$ and UFD $R^q \rightarrow R^r$ of Σ_{FD} , if $R^q \in \text{NDng}(R^p)$ then $R^r \in \text{NDng}(R^p)$.

Proof. Assume by contradiction that $R^r \notin \text{NDng}(R^p)$. Then either $R^r = R^p$ or $R^r \in \text{Dng}(R^p)$. The first case is impossible because of the UFD $R^q \rightarrow R^r$. So we have $R^r \in \text{Dng}(R^p)$. Hence, the UFD $R^r \rightarrow R^p$ is in Σ_{UFD} , so that by transitivity the UFD $R^q \rightarrow R^p$ is in Σ_{UFD} , again contradicting the fact that $R^q \in \text{NDng}(R^p)$. \square

Fix the finite relation-saturated instance I that satisfies Σ_{UFD} , the pssinstance P of I , and the finite Σ_{U} -compliant piecewise realization $PI = (K_1, \dots, K_n)$ of P . Our approach is to construct the desired superinstance I' as $I \sqcup I_1 \sqcup \dots \sqcup I_n$, where the facts of each I_i are constructed from K_i , as we now explain. We call \mathcal{F} the set of the fresh elements (not in $\text{dom}(PI)$) that will be created in the construction, so that we will have $\text{dom}(I') \subseteq \text{dom}(PI) \sqcup \mathcal{F}$.

We consider every $1 \leq i \leq n$. Let R be the relation to which the positions of Π_i belong. If the relation R is not achieved by I and Σ_{UID} , or if Π_i is outer, then we do not create any fact for R , and set $I_i := \emptyset$. Otherwise, as I is relation-saturated, we choose one fact $R(\mathbf{c})$ in I . For every $\mathbf{a} \in K_i \setminus \pi_{\Pi_i}(I)$, we create a fact $F_{\mathbf{a}}^i := R(\mathbf{b})$ in I_i , with b_p defined as follows for every $R^p \in \text{Pos}(\sigma)$:

- If $R^p \in \Pi_i$, take $b_p := a_p$. In other words, the tuple \mathbf{a} is used to fill \mathbf{b} at the positions of Π_i .
- If $R^p \in \text{Dng}(\Pi_i)$, use a fresh element in \mathcal{F} for b_p . In other words, dangerous positions have to be filled with fresh elements (but this is no problem because we will show later that their classes are outer).
- If $R^p \in \text{NDng}(\Pi_i)$ is non-dangerous, take $b_p := c_p$. In other words, we reuse the fact $R(\mathbf{c})$ guaranteed by I being relation-saturated to complete the non-dangerous positions.

We have thus constructed I' , which is clearly a finite superinstance of I . We first show the following claim:

Lemma C.3. *For any $1 \leq i \leq n$ and $\mathbf{a} \in K_i$ for which we create a fact $F_{\mathbf{a}}^i$, for any $R^p \in \Pi_i$, the fact $F_{\mathbf{a}}^i$ is the only fact of I' where a_p occurs at position R^p .*

This claim implies that the facts of I , and all the facts of the I_i for $1 \leq i \leq n$, are pairwise distinct. By this, we mean that we did not try to recreate in I_i a fact that already existed in I , and that we never tried to create the same fact twice in the same I_i or in different I_i .

Proof. Fix $1 \leq i \leq n$ and $\mathbf{a} \in K_i$, and assume that we have created a fact $F_{\mathbf{a}}^i$; fix $R^p \in \Pi_i$.

We first show that we cannot have $a_p \in \pi_{R^p}(I)$. Assuming by contradiction that we do, let F be a witnessing fact. By definition of a piecewise realization we have $\pi_{\Pi_i}(I) \subseteq K_i$, so $\pi_{\Pi_i}(F) \in K_i$. Hence, as PI is Σ_{FD} -compliant, we have $\mathbf{a} = \pi_{\Pi_i}(F)$; but we do not create facts for the tuple $\mathbf{a} \in K_i$ if $\mathbf{a} \in \pi_{\Pi_i}(I)$, which contradicts the fact that we created $F_{\mathbf{a}}^i$.

Second, we show that there cannot be another fact F of $I' \setminus I$ such that $a_p = \pi_{R^p}(F)$. As PI is Σ_{UFD} -compliant, there clearly cannot be such a fact $F_{\mathbf{a}'}^i$ for $\mathbf{a}' \in K_i$, $\mathbf{a} \neq \mathbf{a}'$, with a_p occurring at position R^p of $F_{\mathbf{a}'}^i$. Hence, F is a fact $F_{\mathbf{a}'}^{i'}$ for $i' \neq i$. Now, Π_i and $\Pi_{i'}$ are disjoint as $\leftrightarrow_{\text{FUN}}$ -classes, and thus we cannot have $R^p \in \Pi_{i'}$. So either $R^p \in \text{Dng}(\Pi_{i'})$ and $b_p \in \mathcal{F}$, or $R^p \in \text{NDng}(\Pi_{i'})$ and $b_p \in \pi_{R^p}(I)$. The first case is impossible because elements of \mathcal{F} occur in only one fact, and we showed above that the second case was impossible. This concludes. \square

We now show that I' has the required properties. Let us first show that I' is weakly-sound. Recall the definition:

Definition IV.1. *A superinstance I' of an instance I is **weakly-sound** if the following holds:*

- for any $\mathbf{a} \in \text{dom}(I)$ and $R^p \in \text{Pos}(\sigma)$, if $\mathbf{a} \in \pi_{R^p}(I')$, then either $\mathbf{a} \in \pi_{R^p}(I)$ or $\mathbf{a} \in \text{Wants}(I, R^p)$;
- for any $\mathbf{a} \in \text{dom}(I') \setminus \text{dom}(I)$ and $R^p, S^q \in \text{Pos}(\sigma)$, if $\mathbf{a} \in \pi_{R^p}(I')$ and $\mathbf{a} \in \pi_{S^q}(I')$ then $R^p = S^q$ or $R^p \subseteq S^q$ is in Σ_{UID} .

We begin by checking the first condition. Let $a \in \text{dom}(I)$ and $R^p \in \text{Pos}(\sigma)$ such that $a \in \pi_{R^p}(I')$, and let F be a fact of I' that witnesses it. If F is a fact of I then $a \in \pi_{R^p}(I)$ and a does not witness a violation of weak-soundness. So F is a fact of $I' \setminus I$. Let i be the index of the I_i that contains F , and \mathbf{a} be such that $F = F_a^i$ (this is uniquely defined according to Lemma C.3).

We cannot have $R^p \in \text{Dng}(\Pi_i)$, because we would then have $\Pi_{R^p}(F) \in \mathcal{F}$, contradicting $a \in \text{dom}(I)$. We cannot have $R^p \in \text{NDng}(\Pi_i)$ either, because then $a = \pi_{R^p}(F)$ would imply that $a \in \pi_{R^p}(I)$ which we already excluded. Hence $R^p \in \Pi_i$. Now, by definition of PI being a piecewise realization, as $\mathbf{a} \in K_i$, we know that $a \in \pi_{R^p}(I)$ or $a \in \text{Wants}(P, R^p)$. But we excluded $a \in \pi_{R^p}(I)$ above, and we assumed $a \in \text{dom}(I)$, so $a \in \text{Wants}(P, R^p)$ translates to $a \in \text{Wants}(I, R^p)$. Hence, a does not witness a violation of weak-soundness.

We now check the second condition. Let $a \in \text{dom}(I') \setminus \text{dom}(I)$ and $R^p, S^q \in \text{Pos}(\sigma)$ such that $a \in \pi_{R^p}(I') \cap \pi_{S^q}(I')$. We must show that $R^p = S^q$ or $R^p \subseteq S^q$ is in Σ_{UID} , that is, $R^p \sim_{\text{ID}} S^q$. Now either $a \in \mathcal{F}$, or $a \in \mathcal{H}$. If $a \in \mathcal{F}$, observe that elements of \mathcal{F} occur at only one position in I' . Hence, necessarily $R^p = S^q$ which implies $R^p \sim_{\text{ID}} S^q$, and a does not witness a violation of weak-soundness. Thus, $a \in \mathcal{H}$.

Let F be a fact witnessing that $a \in \pi_{R^p}(I')$, and F' a fact witnessing that $a \in \pi_{S^q}(I')$. As $a \in \mathcal{H}$, necessarily F and F' are facts of $I' \setminus I$, so there are i and i' such that F and F' are respectively facts of I_i and $I_{i'}$. Clearly a cannot occur in F or F' at a position of $\text{Dng}(\Pi_i)$ or $\text{Dng}(\Pi_{i'})$ (they contain elements of \mathcal{F}) or at a position of $\text{NDng}(\Pi_i)$ or $\text{NDng}(\Pi_{i'})$ (they contain elements of $\text{dom}(I)$). Hence, $R^p \in \Pi_i$ and $S^q \in \Pi_{i'}$. Now, as PI is a piecewise realization, as $a \notin \text{dom}(I)$, we conclude that $a \in \text{Wants}(P, R^p)$ and $a \in \text{Wants}(P, S^q)$, and as $a \notin \text{dom}(I)$ this implies that $R^p \in \lambda(a)$ and $S^q \in \lambda(a)$, so that $R^p \sim_{\text{ID}} S^q$, and a does not witness a violation of weak-soundness.

Hence, I' is weakly-sound.

Let us now show that $I' \models \Sigma_{\text{UFD}}$. Assume to the contrary the existence of two facts F and F' that witness a violation of a UFD $\phi : R^p \rightarrow R^q$ of Σ_{UFD} . As $I \models \Sigma_{\text{UFD}}$, we assume without loss of generality that F is a fact of $I' \setminus I$; let $1 \leq i \leq n$ and $\mathbf{a} \in K_i$ be such that $F = F_a^i$. We cannot have $R^p \in \text{Dng}(\Pi_i)$, as then we would have $a_p \in \mathcal{F}$, and elements of \mathcal{F} only occur in a single fact in I' . We cannot have $R^p \in \Pi_i$ either because, by Lemma C.3, F_a^i is the only fact of I' where a_p occurs at position R^p . So $R^p \in \text{NDng}(\Pi_i)$, and by Lemma C.2 we have $R^q \in \text{NDng}(\Pi_i)$ as well. Hence, letting $F'' = R(c)$ be the fact of I used to fill the positions of $\text{NDng}(\Pi_i)$ in F , we know that $a'_p = c_p$ and $a'_q = c_q$. Thus, as this makes it impossible that $F' = F''$, we deduce that F'' and F' also violate ϕ .

Now, either F' is also a fact of I and we have a contradiction because $F'' \in I$ but $I \models \Sigma_{\text{UFD}}$, or it is a fact of $I' \setminus I$ and, by the same process that we applied to F , we can replace it by a fact of I , reaching a contradiction again. This proves that $I' \models \Sigma_{\text{UFD}}$.

Let us last show that $I' \models \Sigma_{\text{UID}}$. Assume to the contrary the existence of a UID $\tau : R^p \subseteq S^q$ of Σ_{UID} and an element $a \in \text{dom}(I')$ such that $a \in \pi_{R^p}(I') \setminus \pi_{S^q}(I')$. Let F be a fact of I' witnessing that $a \in \pi_{R^p}(I')$. Either F is a fact of I or it is a fact of $I' \setminus I$.

For the first case, if F is a fact of I , by definition of PI being a realization, we have $a \in \pi_{R^p}(PI)$. As PI is Σ_{UID} -compliant, we have $a \in \pi_{S^q}(PI)$, and letting \mathbf{a} be the witnessing tuple in K_i where Π_i is the $\leftrightarrow_{\text{FUN}}$ -class of S^q , we know that either $a \in \pi_{S^q}(I)$ or $a \in \pi_{S^q}(F_a^i)$. In the first sub-case there is nothing to show. In the second sub-case it suffices to show that F_a^i was indeed created, and this is the case because τ witnesses that Π_i is inner, and $F \in I$ witnesses that R was achieved in $\text{Chase}(I, \Sigma_{\text{UID}})$, so S must also be because of τ . This concludes the first case.

For the second case, if F is a fact of $I' \setminus I$, write $F = F_a^{i'}$. The existence of $F_a^{i'}$ implies that $\Pi_{i'}$ is inner and R is achieved in $\text{Chase}(I, \Sigma_{\text{UID}})$; hence S is, because of τ . There are three possibilities: $R^p \in \text{NDng}(\Pi_{i'})$, $R^p \in \Pi_{i'}$, or $R^p \in \text{Dng}(\Pi_{i'})$. The first sub-case is $R^p \in \text{NDng}(\Pi_{i'})$; but then we could have picked as witness for $a \in \pi_{R^p}(I')$ the fact $S(c)$ of I used to define the non-dangerous positions, and

we are back to the first case. The second sub-case is $R^p \in \Pi_{i'}$; then we have $a \in \pi_{R^p}(PI)$ by construction, so that as PI is Σ_{UID} -compliant we have $a \in \pi_{S^q}(PI)$, and we conclude as before. The only remaining sub-case is the third sub-case, $R^p \in \text{Dng}(\Pi_{i'})$, so that $a_p \in \mathcal{F}$. Now, as $R^p \in \text{Dng}(\Pi_{i'})$, we know that $R^p \rightarrow R^r$ is in Σ_{UFD} for any position R^r of $\Pi_{i'}$ that occurs in Σ_{UID} (such an R^r exists because $\Pi_{i'}$ is inner). Now, as τ witnesses that R^p occurs in Σ_{UID} , we know by assumption reversible that $R^r \rightarrow R^p$ is in Σ_{UFD} , so that $R^p \in \Pi_{i'}$. But we assumed $R^p \in \text{Dng}(\Pi_{i'})$, a contradiction.

Hence we conclude that $I' \models \Sigma_{\text{UID}}$.

Hence, I' is a finite superinstance of I which is weakly-sound and satisfies Σ_{U} . This concludes the proof.

D. Proofs for Section V: k -Soundness and Reversible UIDs

This section completes the proof of the Acyclic Unary Models Theorem (Theorem III.6) under assumption reversible.

D.1. Proof of Lemma V.2 (ACQs are preserved through k -bounded simulations)

Lemma V.2. *For any instance I and ACQ q of size $\leq n$ such that $I \models q$, if there is an n -bounded simulation from I to I' , then $I' \models q$.*

Fix the instance I . We will prove by induction on n the following stronger claim: for any $n \in \mathbb{N}$, for any ACQ q of size $\leq n$ and any variable x of q , if q has a match in I that maps x to $a \in \text{dom}(I)$, then for any $b \in \text{dom}(I')$ such that $(I, a) \leq_n (I', b)$, q has a match in I' mapping x to b . The base case of $n = 0$ corresponds to queries with no atoms, and it is trivial.

For the induction step, fix $n \in \mathbb{N}$, the query q , the variable x and the match h from q to I that maps x to $a \in \text{dom}(I)$. We define a reachability relation between variables of q as the reflexive and transitive closure of the relation of co-occurring in some atom of q . If this relation consists of a single class, we say that q is **connected**. As we can otherwise rewrite q as a conjunction of strictly smaller queries of ACQ and process all such queries separately using the induction hypothesis, we assume without loss of generality that q is connected.

Let $\mathcal{A} = A_1, \dots, A_m$ be the atoms of q where x occurs (this set of atoms is non-empty, by the connectedness assumption). Because q is an ACQ, each variable y occurring in one of the A_i occurs at most once: once per atom (as the same variable cannot occur multiple times in an atom), and in only one atom (as if y occurs both in A_{i_1} and A_{i_2} then A_{i_1}, y, A_{i_2}, x is a Berge cycle of q). Let Y be the set of the variables occurring in the A_i (not including x).

Because q is acyclic and connected, the other variables of q can be partitioned depending on the variable in Y from which they are reachable without using \mathcal{A} . Hence, we can partition the remaining atoms of q into strictly smaller acyclic subqueries $q_1(y_1, z_1), \dots, q_l(y_l, z_l)$ in ACQ, for $Y = \{y_1, \dots, y_l\}$, where the z_j are pairwise disjoint sets of variables.

Now, let $b \in \text{dom}(I')$ be such that $(I, a) \leq_n (I', b)$. For each atom $A_i = R(x)$ in \mathcal{A} , let $1 \leq p_i \leq |R|$ be the one position such that $x_{p_i} = x$. Consider the fact $F_i = R(a_i)$ that is the image of A_i in I by h . As $(I, a) \leq_n (I', b)$, there exists a fact $F'_i = R(b_i)$ of I' with $b_{p_i} = b$ and with $(I, a_q) \leq_{n-1} (I', b_q)$ for all $1 \leq q \leq |R|$. Consider now each variable $y_j \in Y$ that occurs in A_i , letting $1 \leq q \leq |R|$ be the one position such that $x_q = y_j$, and let $q_j(y_j, z_j)$ be the subquery corresponding to y_j . We know that $(I, a_q) \leq_{|q_j|} (I', b_q)$, and that q_j has a match in I that maps y_j to a_q (namely, the restriction h_j of the match h to the subquery q_j) so that, by the induction hypothesis, q_j has a match h'_j in I' where y_j is

matched to b_q . Now, we can assemble the F'_i and all the matches h'_j thus obtained, because the z_j are pairwise disjoint, yielding a match h' of q in I' where x is matched to b . This concludes the induction step.

Hence, the stronger claim is proven by induction. It remains to observe that it implies the desired claim. Indeed, if $I \models q$ and there is a n -bounded simulation sim from I to I' , choose any variable x in q (if q has no variables, the result is vacuous), consider any match of q in I matching x to a , use sim to define $b := \text{sim}(a)$, and deduce the existence of a match of q in I' (matching x to b) using the claim that we have shown by induction.

D.2. Proof of Lemma V.5 (AFactCl is finite)

Lemma V.5. *For any initial instance I_0 , set Σ_{UID} of UUIDs, and $k \in \mathbb{N}$, AFactCl is finite.*

We first show that \simeq_k has only a finite number of equivalence classes on $\text{Chase}(I_0, \Sigma_{\text{UID}})$. Indeed, for any element $a \in \text{dom}(\text{Chase}(I_0, \Sigma_{\text{UID}}))$, by the Unique Witness Property, the number of facts in which a occurs is bounded by a constant depending only on I_0 and Σ_{UID} . Hence, there is a constant M depending only on I_0 , Σ_{UID} , and k , so that, for any element $d \in \text{dom}(\text{Chase}(I_0, \Sigma_{\text{UID}}))$, the number of elements of $\text{dom}(\text{Chase}(I_0, \Sigma_{\text{UID}}))$ which are relevant to determine the \simeq_k -class of d (that is, the elements whose distance to d in the Gaifman graph of $\text{Chase}(I_0, \Sigma_{\text{UID}})$ is $\leq k$) is bounded by M .

This clearly implies that AFactCl is finite, because the number of m -tuples of equivalence classes of \simeq_k that occur in $\text{Chase}(I_0, \Sigma_{\text{UID}})$ is then finite for any $m \leq \max_{R \in \sigma} |R|$, and $\text{Pos}(\sigma)$ is finite.

D.3. Proof of the Fact-Saturated Solutions Lemma (Lemma V.6)

Lemma V.6 (Fact-saturated solutions). *The result I of performing sufficiently many chase rounds on I_0 is such that $J_0 = (I, \text{id})$ is a fact-saturated aligned superinstance of I_0 .*

For every $D \in \text{AFactCl}$, let $n_D \in \mathbb{N}$ be such that D is achieved by a fact of $\text{Chase}(I_0, \Sigma_{\text{UID}})$ created at round n_D . As AFactCl is finite, $n := \max_{D \in \text{AFactCl}} n_D$ is finite. Hence, all classes of AFactCl are achieved after n chase rounds on I_0 .

Consider now I'_0 obtained from the aligned superinstance I_0 by n rounds of the UUID chase, and $J_0 = (I'_0, \Sigma_{\text{UID}})$. It is clear that for any $D \in \text{AFactCl}$, there is an achiever $F = R(\mathbf{b})$ of D in I'_0 . Hence, the corresponding fact in J_0 is an achiever of D in J_0 .

D.4. Proof of the Fact-Thrifty Chase Steps Lemma (Lemma V.9)

We first prove the following lemma, which we will use to justify that we can extend aligned instances.

Lemma D.1. *Let $n \in \mathbb{N}$. Let I_1 and I be instances and sim be a n -bounded simulation from I_1 to I . Let I_2 be a superinstance of I_1 defined by adding one fact $F_n = R(\mathbf{a})$ to I_1 , and let sim' be a mapping from I_2 to I such that $\text{sim}'|_{I_1} = \text{sim}$. Assume there is a fact $F_w = R(\mathbf{b})$ in I such that, for all $R^i \in \text{Pos}(R)$, $\text{sim}'(a_i) \simeq_n b_i$. Then sim' is a n -bounded simulation from I_2 to I .*

Proof. We prove the claim by induction on n . The base case of $n = 0$ is immediate.

Let $n > 0$, assume that the claim holds for $n - 1$, and show that it holds for n . As sim is a n -bounded simulation, it is a $(n - 1)$ -bounded simulation, so we know by the induction hypothesis that sim' is a $(n - 1)$ -bounded simulation.

Let us now show that it is a n -bounded simulation. Let $a \in \text{dom}(I_2)$ be an element and show that $(I_2, a) \leq_n (I, \text{sim}'(a))$. To do this, choose $F = S(\mathbf{a})$ a fact of I_2 with $a_p = a$ for some p , and show that there exists a fact $F' = S(\mathbf{a}')$ of I with $a'_p = \text{sim}'(a_p)$ and $(I_2, a_q) \leq_{n-1} (I, a'_q)$ for all $S^q \in \text{Pos}(S)$.

The first possibility is that F is the new fact $F_n = R(\mathbf{a})$. In this case, as we have $(I, b_p) \leq_n (I, \text{sim}'(a_p))$, considering F_w , we deduce the existence of a fact $F'_w = R(\mathbf{c})$ in I such that $c_p = \text{sim}'(a_p)$ and $(I, b_q) \leq_{n-1} (I, c_q)$ for all $1 \leq q \leq |R|$. We take $F' = F'_w$. By construction we have $c_p = \text{sim}'(a_p)$. Fixing $1 \leq q \leq |R|$, to show that $(I_2, a_q) \leq_{n-1} (I, c_q)$, we use the fact that sim' is an $(n-1)$ -bounded simulation to deduce that $(I_2, a_q) \leq_{n-1} (I, \text{sim}'(a_q))$. Now, we have $(I, \text{sim}'(a_q)) \leq_{n-1} (I, b_q)$, and as we explained we have $(I, b_q) \leq_{n-1} (I, c_q)$, so we conclude by transitivity.

If F is another fact, then it is a fact of I_1 , so its elements are in $\text{dom}(I_1)$, and as sim' coincides with sim on such elements, we conclude because sim is a n -bounded simulation. \square

We then prove the main result:

Lemma V.9 (Fact-thrifty chase steps). *For any fact-saturated aligned superinstance J , the result J' of a fact-thrifty chase step on J is indeed a well-defined aligned superinstance where the former active fact F_a is no longer active.*

We first observe that fact-thrifty chase steps are well-defined because a suitable $F_t = S(\mathbf{c})$ always exists, as J is fact-saturated. It is immediate that J' is finite.

It is immediate that, letting $J' = (I', \text{sim}')$ be the result of the process, I' is still a superinstance of I_0 , and the previously active fact F_a is no longer active in I' . To show that sim' is still a k -bounded simulation, use Lemma D.1 with $F_n = S(\mathbf{b})$ and $F_w = S(\mathbf{b}')$. The fact that sim' is the identity on I_0 is immediate because $\text{sim}'|_{I_0} = \text{sim}|_{I_0}$.

We now show that J' satisfies Σ_{UFD} , using the fact that J does. Indeed, any violation of Σ_{UFD} in J' would have to include the one new fact $F_n = S(\mathbf{b})$. By way of contradiction, let $\phi : S^l \rightarrow S^r$ be a violated UFD in Σ_{UFD} and let $\{F, F_n\}$ be a violation, where $F = S(\mathbf{d})$ is some fact of I' . It is clear that we cannot have $d_q = b_q$, as otherwise this would contradict the fact that F_a was an active fact. Hence, by construction of the new fact F_n , we can only have $b_i = d_i$ if $S^i \in \text{NDng}(S^q)$. As $\{F, F_n\}$ violates ϕ , this implies that $S^l \in \text{NDng}(S^q)$, so that, by Lemma C.2, $S^r \in \text{NDng}(S^q)$. Now, observe that we have $\pi_{\text{NDng}(S^q)}(F_n) = \pi_{\text{NDng}(S^q)}(F_r)$, with F_r the fact used to fill the non-dangerous position in the definition of fact-thrifty chase steps. Now, we cannot have $F = F_r$ because they must disagree on S^r , so that $\{F, F_r\}$ also witnesses a violation of ϕ in J . This contradicts our assumption that $J \models \Sigma_{\text{UFD}}$.

We must now check the last part of the definition of aligned superinstances, which only needs to be verified for the fresh elements: for $S^r \neq S^q$, if b_r is fresh, then it occurs in J' at the position where $\text{sim}(b_r)$ was introduced in $\text{Chase}(I_0, \Sigma_{\text{UID}})$. For this, it suffices to show that b'_q was the exported element of F_w . In this case, as $\text{sim}(b_r) = b'_r$, we will know that b'_r was introduced at position S^r in F_w in $\text{Chase}(I_0, \Sigma_{\text{UID}})$, so the condition is respected. We make this a separate lemma:

Lemma D.2. *Let J be an aligned superinstance of I_0 and consider the application of a thrifty chase step for a UID $\tau : R^p \subseteq S^q$. Consider the chase witness $F_w = S(\mathbf{b}')$. Then b'_q is the exported element of F_w .*

Using this lemma, it is also clear that $\text{sim}'|_{I' \setminus I_0}$ maps to $\text{Chase}(I_0, \Sigma_{\text{UID}}) \setminus I_0$, which is the last thing we had to verify. Indeed, for all fresh elements $b_r \in \text{dom}(I') \setminus \text{dom}(I)$ (with $S^r \neq S^q$), which are clearly not in I_0 , we have fixed $\text{sim}'(b_r)$ to be b'_r , which by the lemma is introduced in F_w so it cannot be an element of I_0 ; hence it is indeed an element of $\text{Chase}(I_0, \Sigma_{\text{UID}}) \setminus I_0$.

We conclude by proving Lemma D.2:

Proof. Let $F_a = R(a)$ be the active fact in J , $F_b = S(b)$ be the new fact of J' , and $\tau : R^p \subseteq S^q$ be the UID, so $a_p = b_q$ is the exported element of this chase step. Assume by way of contradiction that b'_q was not the exported element in F_w , so that it was introduced in F_w . In this case, as $\text{sim}(a_p) = \text{sim}(b_q) = b'_q$, by the last part of the definition of aligned superinstances, we have $a_p \in \pi_{S^q}(J)$, which contradicts the fact that $a_p \in \text{Wants}(J, \tau)$. Hence, we have proved by contradiction that b'_q was the exported element in F_w . \square

D.5. Proof of the Fact-Thrifty Completion Proposition (Proposition V.10)

Proposition V.10 (Fact-thrifty completion). *Under assumption reversible, for any fact-saturated aligned superinstance J of I_0 , we can expand J by fact-thrifty chase steps to a fact-saturated aligned superinstance J' of I_0 that satisfies Σ_{UID} .*

There are two steps to the proof. The first one is to apply initial chasing by fresh fact-thrifty chase steps to ensure a certain property, *k-reversibility*. The second one is to use fact-thrifty chase steps to satisfy Σ_{UID} , using the constructions of Section IV.

We start with the first step. We consider a forest structure on the facts of $\text{Chase}(I_0, \Sigma_{\text{UID}})$: the facts of I_0 are the roots, and the parent of a fact F not in I_0 is the fact F' that was the active fact for which F was created, so that F' and F share the exported element of F . For $a \in \text{dom}(\text{Chase}(I_0, \Sigma_{\text{UID}}))$, if a was introduced at position S^r of an S -fact $F = S(a)$ created by applying the UID $\tau : R^p \subseteq S^q$ (with $S^q \neq S^r$) to its parent fact F' , we call τ the *last* UID of a . The last two UIDs of a are (τ, τ') where τ' is the last UID of the exported element a_q of F (which was introduced in F'). For $n \in \mathbb{N}$, we define the last n UIDs in the same way, for elements of $\text{Chase}(I_0, \Sigma_{\text{UID}})$ introduced after sufficiently many rounds. We say that a is *n-reversible* if its last n UIDs are reversible.

We accordingly define the notion of *n-reversible aligned superinstance*, which requires that elements where a UID is violated are mapped by sim to a *n-reversible* element in the chase. Recall that, for any position R^p , we write $[R^p]_{\text{ID}}$ the \sim_{ID} -class of R^p .

Definition D.3. *An aligned superinstance J of I_0 is **n-reversible** if for any position S^q and $a \in \text{Wants}(J, S^q)$, $\text{sim}(a)$ is a *n-reversible* element of $\text{Chase}(I_0, \Sigma_{\text{UID}})$ introduced at a position of $[S^q]_{\text{ID}}$ in $\text{Chase}(I_0, \Sigma_{\text{UID}})$.*

The first step of the proof of Proposition V.10 is to perform $k + 1$ fresh fact-thrifty chase rounds on the input fact-saturated aligned superinstance J , to ensure that the result J' is *k-reversible* for Σ_{UID} :

Proposition D.4 (Ensuring *n-reversibility*). *For any $n \in \mathbb{N}$, applying $n + 1$ fresh fact-thrifty chase rounds on a fact-saturated aligned superinstance J by the UIDs of Σ_{UID} yields a fact-saturated aligned superinstance J' that is *n-reversible* for Σ_{UID} .*

This proposition is proved in Appendix D.6.

The second step of the proof is simply to apply the following lemma to J' .

Lemma D.5 (Guided chase). *For any fact-saturated *k-reversible* aligned superinstance $J = (I, \text{sim})$ of I_0 , we can build by fact-thrifty chase steps an aligned superinstance $J' = (I', \text{sim}')$ of I_0 such that $I \subseteq I'$, $\text{sim}'|_I = \text{sim}$, and J' satisfies Σ_{UID} .*

The lemma is proved in Appendix D.7. It uses the constructions of Section IV, and relies on an independent result about the UID chase, the Chase Locality Theorem (Theorem V.11), proved in Appendix D.8. Clearly, applying the Guided Chase Lemma to J' concludes the proof of the Fact-Thrifty Completion Proposition.

D.6. Proof of Proposition “Ensuring n -reversibility” (Proposition D.4)

We first make the following easy observation:

Lemma D.6. *Let J be an aligned superinstance, and J' be the result of applying one chase round to J with fresh fact-thrifty chase steps. Let $a \in \text{Wants}(J', \tau)$ for any UID τ . Then we have $a \in \text{dom}(J') \setminus \text{dom}(J)$, and a occurs in a single fact F (which is an active fact for τ).*

Proof. For the first part of the claim, let us assume by way of contradiction that $a \in \text{dom}(J)$. Note that, by definition of chase rounds, we cannot have $a \in \text{Wants}(J, \tau)$, otherwise we could not have $a \in \text{Wants}(J', \tau)$. Hence, if we have $a \in \text{dom}(J)$ but $a \notin \text{Wants}(J, \tau)$, any active fact F witnessing $a \in \text{Wants}(J, \tau)$ must be in J' .

Now, by definition of fact-thrifty chase steps, if $a \notin \text{dom}(J)$, there are two possibilities. Either a was the exported element in F , or it was an element reused at a non-dangerous position. The first case is impossible: because Σ_{UID} is transitively closed, the new facts created in J' cannot make new UIDs applicable to old elements of J . The second case is also impossible: elements reused at non-dangerous positions already occurred at the same position in J , so this cannot make new UIDs applicable to them. This proves the first claim.

The second part of the claim is by observing that elements created in J' occur in a single fact, by definition of chase rounds, and by definition of fresh fact-thrifty chase steps (elements in new facts are either in $\text{dom}(J)$ or are fresh). So the one fact where a occurs must be the active fact witnessing that $a \in \text{Wants}(J', \tau)$. \square

We then show the following simple lemma about n -reversibility:

Lemma D.7. *Let $n \in \mathbb{N}$, let J be a n -reversible aligned superinstance of I_0 and let $F_n = S(\mathbf{b})$ be a new fact obtained by applying a thrifty chase step to J . For all $S^r \in \text{Pos}(S)$, such that $b_r \notin \text{dom}(J)$, $\text{sim}(b_r)$ is $(n+1)$ -reversible and introduced at position S^r in $\text{Chase}(I_0, \Sigma_{\text{UID}})$.*

Proof. Let F_a be the active fact, F_w be the chase witness, and $\tau : R^p \subseteq S^q$ be the UID for this chase step. By Lemma D.2 we know that b'_q is the exported element of F_w . Hence, for all $S^r \in \text{Pos}(S) \setminus \{S^q\}$, b'_r is $(n+1)$ -reversible and introduced at position S^r . Now, for all $S^r \in \text{Pos}(S)$ such that b_r is fresh in F_n , we have $\text{sim}(b_r) = b'_r$, so the result follows. \square

We now prove the main result:

Proposition D.4 (Ensuring n -reversibility). *For any $n \in \mathbb{N}$, applying $n+1$ fresh fact-thrifty chase rounds on a fact-saturated aligned superinstance J by the UIDs of Σ_{UID} yields a fact-saturated aligned superinstance J' that is n -reversible for Σ_{UID} .*

Fix the aligned superinstance $J = (I, \text{sim})$. We prove the result by induction on n . For the base case $n = 0$, letting J' be the result of applying one chase round to J , we need only show that for any position S^q and $a \in \text{Wants}(J', S^q)$, $\text{sim}(a)$ was introduced at a position of $[S^q]_{\text{ID}}$ in $\text{Chase}(I_0, \Sigma_{\text{UID}})$. By Lemma D.6, a occurs in a single fact F at some position R^p (so that, using assumption reversible, $R^p \sim_{\text{ID}} S^q$), and we have $a \in \text{dom}(J') \setminus \text{dom}(J)$, so it was created by the application of a thrifty chase step to J . By Lemma D.7, we conclude that $\text{sim}(a)$ was introduced at position R^p in $\text{Chase}(I_0, \Sigma_{\text{UID}})$, which implies the desired claim.

For the induction, fix $n > 0$ and assume that the result is true for $n-1$. Let $J' = (I', \text{sim}')$ be the result of applying $(n-1)+1$ chase rounds to J . By induction hypothesis, J is $(n-1)$ -reversible. We want to show that $J'' = (I'', \text{sim}'')$ obtained by applying one more chase round to J' is n -reversible. This is shown

exactly as in the base case, except that, when applying Lemma D.7, we use the $(n - 1)$ -reversibility of J to deduce the n -reversibility of the element under consideration.

This proves the desired claim by induction. Note that we have relied implicitly on the Fact-Thrifty Chase Steps Lemma (Lemma V.9) to justify that the result of chase rounds by fact-thrifty chase steps are indeed aligned superinstances; it is immediate that fact-saturation is preserved.

D.7. Proof of the Guided Chase Lemma (Lemma D.5)

Recall that the $\leftrightarrow_{\text{FUN}}$ -classes of $\text{Pos}(\sigma)$ are numbered Π_1, \dots, Π_n . Recall the notion of inner and outer $\leftrightarrow_{\text{FUN}}$ -classes (Definition C.1), and the notion of piecewise realization (Definition IV.14). We define:

Definition D.8. A superinstance I' of the instance I follows the piecewise realization $PI = (K_1, \dots, K_n)$ if for every inner $\leftrightarrow_{\text{FUN}}$ -class Π_i , we have $\pi_{\Pi_i}(I') \subseteq K_i$.

We show the main claim:

Lemma D.5 (Guided chase). For any fact-saturated k -reversible aligned superinstance $J = (I, \text{sim})$ of I_0 , we can build by fact-thrifty chase steps an aligned superinstance $J' = (I', \text{sim}')$ of I_0 such that $I \subseteq I'$, $\text{sim}'|_I = \text{sim}$, and J' satisfies Σ_{UID} .

Fix the fact-saturated k -reversible aligned superinstance $J = (I, \text{sim})$ of I_0 . Let $P = (I, \mathcal{H}, \lambda)$ be a balanced pssinstance of J obtained by the Balancing Lemma (Lemma IV.9) and let $PI = (K_1, \dots, K_n)$ be a finite Σ_{U} -compliant piecewise realization of P obtained by the Realizations Lemma (Lemma IV.16).

We will prove the result by satisfying UID violations in J with fact-thrifty chase steps using the piecewise realization PI , yielding a finite aligned superinstance $J_f = (I_f, \text{sim}_f)$ such that $I \subseteq I_f$, the restriction of sim_f to I is sim , J_f satisfies Σ_{UID} , and I_f follows PI . The process is a variant of Lemma “Using realizations to get completions” (Lemma IV.21).

We call $J' = (I', \text{sim}')$ the current state of our superinstance, starting at $J' := J$. We will perform fact-thrifty chase steps on J' . We call \mathcal{F} the set of all fresh elements (not in $\text{dom}(P)$) that we will introduce (only in outer classes) during the chase steps. It is immediate that our construction will maintain the following:

fsat: J' is a fact-saturated aligned superinstance of I_0 (this uses Lemma V.9);

sub: $I \subseteq I'$;

sim: $\text{sim}'|_{\text{dom}(I)} = \text{sim}$.

Further, we will additionally maintain the following invariants:

fw: I' follows PI ;

krev: J' is k -reversible;

out: elements of outer classes are only in \mathcal{F} or in $\text{dom}(I)$.

We now describe formally how we apply each fact-thrifty chase step. Choose an element $a \in \text{Wants}(J', \tau)$ to which some UID $\tau : R^p \subseteq S^q$ is applicable. Let $F_a = R(\mathbf{a})$ be the active fact, with $a = a_p$. The UID τ witnesses that the $\leftrightarrow_{\text{FUN}}$ -classes Π_i and $\Pi_{i'}$, of R^p and S^q respectively, are inner, so

by invariant fw we have $a \in \pi_{R^p}(PI)$. As PI is Σ_{UID} -compliant, we must have $a \in \pi_{S^q}(PI)$, and there is a $|\Pi_{I'}|$ -tuple $t \in K_{I'}$ such that $t_q = a$.

We choose a fact $F_r = S(c)$ of J that achieves the fact class of the chase witness F_w (this is possible by invariant fsat), and create a new fact $F_n = S(b)$ with the fact-thrifty chase step defined as follows:

- For the exported position S^q , we set $b_q := a_p$.
- For any $S^r \in \Pi_{I'}$, noting that necessarily $S^r \in \text{Dng}(S^q)$, we set $b_r := t_r$.
- For any position $S^r \in \text{Dng}(S^q) \setminus \Pi_{I'}$, we take b_r to be a fresh element from \mathcal{F} .
- For any position $S^r \in \text{NDng}(S^q)$, we set $b_r := c_r$.

We must verify that this satisfies the conditions of thrifty chase steps. The fact that $b_r \in \pi_{S^r}(J')$ for $S^r \in \text{NDng}(S^q)$ is immediate by definition of F_r . We now show the two other points.

First, we show that $b_r \notin \pi_{S^r}(J')$ for $S^r \in \text{Dng}(S^q)$. Obviously this needs only to be checked for $S^r \in \Pi_{I'}$ (as the other b_r are always fresh). Assume to the contrary that $t_r \in \pi_{S^r}(J')$, and let $F = S(d)$ be a witnessing fact. As $\Pi_{I'}$ is inner, by invariant fw, we deduce that $\pi_{\Pi_{I'}}(d) \in \pi_{\Pi_{I'}}(PI)$. Now, as $d_r = t_r$ and PI is Σ_{UFD} -compliant, we deduce that $d = t$, so that F witnesses that d_q is in $\pi_{S^q}(J')$. As we have $d_q = t_q = a$, this contradicts the applicability of τ to a . Hence, the claim is proven.

Second, we check that reused elements have the right sim-image. This is the case by definition of fact-thrifty chase steps for the non-dangerous positions, so again we need only check this for elements at a position $S^r \in \Pi_{I'}$, and only if they are not fresh. We start by showing that, for such S^r , we have $b_r \in \text{Wants}(J', S^r)$.

Indeed, we have $b_r = t_r$ which is in $\pi_{S^r}(PI)$, and we cannot have $t \in \pi_{\Pi_{I'}}(J')$, as otherwise this would contradict the applicability of τ to a ; so in particular, by invariant sub, we cannot have $t \in \pi_{\Pi_{I'}}(I)$. Thus, by definition of a piecewise realization, we have $t_r \in \text{Wants}(P, S^r)$. Recalling that we have $t_r \in \text{dom}(J')$, we show that this implies $t_r \in \text{Wants}(J', S^r)$. Recalling the definition of $t_r \in \text{Wants}(P, S^r)$, we distinguish two subcases: (1.) $t_r \in \text{dom}(J)$ and $t_r \in \text{Wants}(J, S^r)$, or (2.) $t_r \in \mathcal{H}$ and $S^r \in \lambda(t_r)$.

In the subcase (1.) $t_r \in \text{dom}(J)$ and $t_r \in \text{Wants}(J, S^r)$, we remember that in the first point we showed that $t_r \notin \pi_{S^r}(J')$. So we still have $t_r \in \text{Wants}(J', S^r)$, which is what we claimed.

In the subcase (2.) $t_r \in \mathcal{H}$ and $S^r \in \lambda(t_r)$, consider a fact F' of J' witnessing $t_r \in \text{dom}(J')$, where t_r occurs at a position T^l ; let $\Pi_{I''}$ be the $\leftrightarrow_{\text{FUN}}$ -class of T^l . As $t_r \in \mathcal{H}$, by invariant out, $\Pi_{I''}$ is inner, so by invariant fw there is a tuple t' of $K_{I''}$ such that $t'_l = t_r$. Now, as $t_r \in \mathcal{H}$, by definition of piecewise realizations, we have $T^l \in \lambda(t_r)$. Hence, either the UID $\tau' : T^l \subseteq S^r$ is in Σ_{UID} or we have $T^l = S^r$. As $t_r \in \pi_{T^l}(J')$ and we have shown in the first point that $t_r \notin \pi_{S^r}(J')$, we know that $T^l \neq S^r$, so τ' is in Σ_{UID} . Hence, as F' witnesses that $t_r \in \pi_{T^l}(J')$, and as $t_r \notin \pi_{S^r}(J')$, we have $t_r \in \text{Wants}(J', S^r)$, as we claimed.

Hence, we know that $b_r = t_r$ is in $\text{Wants}(J', S^r)$ in either subcase. By invariant krev, this implies that $\text{sim}(b_r)$ is a k -reversible element of $\text{Chase}(I_0, \Sigma_{\text{UID}})$ introduced at a position of $[S^r]_{\text{ID}}$. By Lemma D.7, we know that the sim-image b'_r of a fresh element at position S^r would be k -reversible and introduced at position S^r . Hence, by the Chase Locality Theorem (Theorem V.11), we have $\text{sim}(b_r) \simeq_k b'_r$, so the condition is satisfied. This proves that, indeed, we can perform the fact-thrifty chase step that we described.

We now check that the invariants are preserved. We first observe that for any $S^r \in \text{Dng}(S^q) \setminus \Pi_{I'}$, the $\leftrightarrow_{\text{FUN}}$ -class of S^r is outer. Indeed, if S^r occurred in Σ_{UID} , as S^q does because of τ , we know by assumption reversible that, as the UFD $S^r \rightarrow S^q$ is in Σ_{UFD} by dangerousness of S^r , the UFD $S^q \rightarrow S^r$ also should, but then we would have $S^r \leftrightarrow_{\text{FUN}} S^q$, so $S^r \in \Pi_{I'}$, a contradiction. Hence, the $\leftrightarrow_{\text{FUN}}$ -class of S^r is indeed outer.

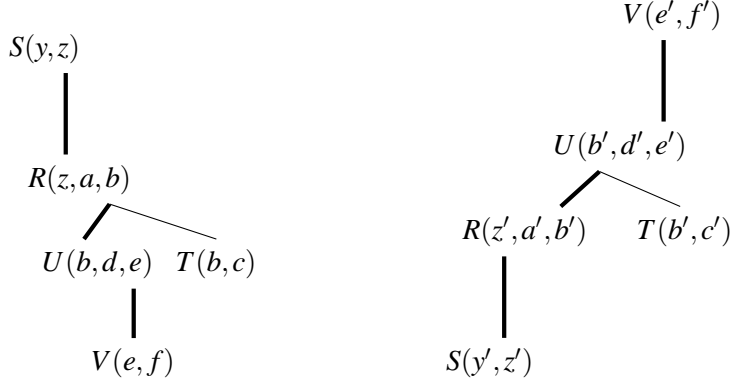


Figure 1: Chase locality example. Elements b and b' are 1-reversible and introduced at positions R^3 and U^1 . Reversible UIDs are represented by thick edges.

Now, invariant fw is preserved because, by the above observation, the new fact F_n is defined on the inner classes either following t or following an existing fact of J' . Invariant krev is preserved by Lemma D.7 for the fresh elements, or by krev on the previous state J' for the existing elements. Invariant out is preserved because the only elements of F_n that are not in \mathcal{F} or in $\text{dom}(I)$ are those of $\Pi_{i'}$, which is inner. This shows that the invariant is preserved by the fact-thrifty chase step.

We perform fact-thrifty chase steps until no violations of Σ_{UID} remain: invariant fw guarantees that we terminate. Indeed, PI is finite, the domain of the resulting instance is bounded by that of PI for all inner classes, and new elements created in outer classes cannot create violations of Σ_{UID} or cause the creation of further elements, by definition of their class being outer. Hence, the result of the process is finite, and it satisfies Σ_{UID} because no violations remain. This concludes the proof.

D.8. Proof of the Chase Locality Theorem (Theorem V.11)

We give an equivalent rephrasing of the Chase Locality Theorem (Theorem V.11) using the notion of n -reversible elements (Definition D.3):

Theorem D.9 (Chase locality theorem). *For any instance I_0 , transitively closed set of UIDs Σ_{UID} , and $n \in \mathbb{N}$, for any two elements a and b respectively introduced at positions R^p and S^q in $\text{Chase}(I_0, \Sigma_{\text{UID}})$ such that $R^p \sim_{\text{ID}} S^q$, if a and b are n -reversible then $a \simeq_n b$.*

Note that this result is for an arbitrary set of UIDs and FDs, not relying on any finite closure properties, or on assumption reversible . (It only assumes that the last n dependencies used to create a and b were reversible.) However we still assume that Σ_{UID} is transitively closed.

Figure 1 illustrates the result in a simple situation. The intuition is the following: n -reversible elements in the chase have the same neighborhoods up to distance n , no matter their exact histories, as long as they were introduced in \sim_{ID} -equivalent positions: intuitively, the facts that go “downwards” in the neighborhood of a in the forest structure can be matched to facts in the neighborhood of b because they are required by Σ_{UID} , and the facts “upwards” are also matched up to distance n because of the reverses of the UIDs used along this chain.

To prove the theorem, fix the instance I_0 and the set Σ_{UID} of UIDs. We first show the following easy lemma:

Lemma D.10. *For any $n > 0$ and position R^p , for any two elements a, b of $\text{Chase}(I_0, \Sigma_{\text{UID}})$ introduced at position R^p in two facts F_a and F_b , letting a' and b' be the exported elements of F_a and F_b , if $a' \simeq_{n-1} b'$, then $a \simeq_n b$.*

Proof. By symmetry, it suffices to show that $a \leq_n b$. We proceed by induction on n .

For the base case $n = 1$, observe that, for every fact F of $\text{Chase}(I_0, \Sigma_{\text{UID}})$ where a occurs at some position S^q , there are only two cases. Either $F = F_a$, so we can pick F_b as the representative fact, or the UID $R^p \subseteq S^q$ is in Σ_{UID} so we can pick a corresponding fact for b by definition of the chase.

For the induction step, we proceed in the same way. If $F = F_a$, we pick F_b and use either the hypothesis on a' and b' or the induction hypothesis (for other elements of F_a and F_b) to justify that F_b is a suitable witness. Otherwise, we pick the corresponding fact for b which must exist by definition of the chase, and apply the induction hypothesis to the other elements of the fact to conclude. \square

We now prove the Chase Locality Theorem. Recall the definition of \sim_{ID} (Definition IV.6). However, note that, as we no longer make assumption reversible, while \sim_{ID} is still an equivalence relation, it is no longer the case that all UIDs of Σ_{UID} are reflected in \sim_{ID} : the UID $R^p \subseteq S^q$ may be in Σ_{UID} even though $R^p \not\sim_{\text{ID}} S^q$ if $S^q \subseteq R^p$ is not in Σ_{UID} .

We prove by induction on n the main claim: for any positions R^p and S^q such that $R^p \sim_{\text{ID}} S^q$, for any two n -reversible elements a and b respectively introduced at positions R^p and S^q , we have $a \simeq_n b$. By symmetry it suffices to show that $(\text{Chase}(I_0, \Sigma_{\text{UID}}), a) \leq_n (\text{Chase}(I_0, \Sigma_{\text{UID}}), b)$.

The base case of $n = 0$ is immediate.

For the induction step, fix $n > 0$, and assume that the result holds for $n - 1$. Fix R^p and S^q , and let a, b be two n -reversible elements introduced respectively at R^p and S^q in facts F_a and F_b . Note that by the induction hypothesis we already know that $(\text{Chase}(I_0, \Sigma_{\text{UID}}), a) \leq_{n-1} (\text{Chase}(I_0, \Sigma_{\text{UID}}), b)$; we must show that this holds for n .

First, observe that, as a and b are n -reversible with $n > 0$, they are not elements of I_0 . Hence, by definition of the chase, for each one of them, the following is true: for each fact of the chase where the element occurs, it only occurs at one position, and all other elements co-occurring with it in a fact of the chase occur only at one position in only one of these facts. Thus, to prove the claim, it suffices to construct a mapping ϕ from the set $N_1(a)$ of the facts of $\text{Chase}(I_0, \Sigma_{\text{UID}})$ where a occurs, to the set $N_1(b)$ of the facts where b occurs, such that the following holds: for every fact $F = T(a)$ of $N_1(a)$, letting T^c be the position of F such that $a_c = a$ (there is only one such position by construction of the chase), b occurs at position T^c in $\phi(F) = T(b)$, and for every i , $a_i \leq_{n-1} b_i$.

By construction of the chase (using the Unique Witness Property), $N_1(a)$ consists of exactly the following facts:

- The fact $F_a = R(a)$, where $a_d = a'$ is the exported element (for a certain d), $a_p = a$ was introduced at R^p in F_a , and for $i \notin \{p, d\}$, a_i was introduced at R^i in F_a
- For every UID $\tau : R^p \subseteq V^g$ of Σ_{UID} , a V -fact F_a^τ where all elements were introduced in this fact except the one at position V^g which is a .

A similar characterization holds for b , with the analogous notation. We construct the mapping ϕ as follows:

- If $R^p = S^q$ then set $\phi(F_a) = F_b$; otherwise, as $\tau : S^q \subseteq R^p$ is in Σ_{UID} , set $\phi(F_a)$ to be the fact F_b^τ .
- For every UID $\tau : R^p \subseteq V^g$ of Σ_{UID} , as $R^p \sim_{\text{ID}} S^q$, by transitivity, either $S^q = V^g$ or the UID $\tau' : S^q \subseteq V^g$ is in Σ_{UID} . In the first case, set $\phi(F_a^\tau) = F_b$. In the second case, set $\phi(F_a^\tau) = F_b^{\tau'}$.

We must now show that ϕ satisfies the required conditions. First, verify that indeed, by construction, whenever a occurs at position T^c in F then b occurs at position T^c in $\phi(F)$. Second, fix $F \in N_1(a)$, write $F = T(a)$ and $\phi(F) = T(b)$, with $a_c = a$ and $b_c = b$ for some c , and show that $a_i \leq_{n-1} b_i$ for all $T^i \in \text{Pos}(T)$. If $n = 1$ there is nothing to show and we are done, so we assume $n \geq 2$. If $i = c$ then the claim is immediate by the induction hypothesis; otherwise, we distinguish two cases:

1. $F = F_a$ (so that $T = R$ and $c = p$), or $F = F_a^\tau$ such that the UID $\tau : R^p \subseteq T^c$ is reversible. In this case, by construction, either $\phi(F) = F_b$ or $\phi(F) = F_b^{\tau'}$ for $\tau' : S^q \subseteq T^c$; τ' is then reversible, because $R^p \sim_{\text{ID}} S^q$ and $R^p \sim_{\text{ID}} T^c$.

We show that for all $1 \leq i \leq |T|$, $i \neq c$, a_i is $(n-1)$ -reversible and was introduced in $\text{Chase}(I_0, \Sigma_{\text{UID}})$ at a position in the \sim_{ID} -class of T^i . Once we have proved this, by symmetry we can show the same for all b_i , so that we can conclude that $a_i \leq_{n-1} b_i$ by induction hypothesis. To see why the claim holds, we distinguish two subcases. Either a_i was introduced in F , or we have $F = F_a$, $i = d$ and a_i is the exported element for a .

In the first subcase, a_i was created by applying the reversible UID τ and the exported element a is n -reversible, so a_i is $(n-1)$ -reversible (in fact it is $(n+1)$ -reversible), and is introduced at position T^i . In the second subcase, a_i is the exported element used to create a , which is n -reversible, so a_i is $(n-1)$ -reversible; and as $n \geq 2$, the last dependency applied to create a_i is reversible, so that a_i was introduced at a position in the same \sim_{ID} -class as T^i . Hence, we have proved the desired claim in the first case.

2. $F = F_a^\tau$ such that $\tau : R^p \subseteq T^c$ is not reversible. In this case, we cannot have $T^c = S^q$ (because we have $R^p \sim_{\text{ID}} S^q$), so that $\phi(F) = F_b^\tau$, and all a_i for $i \neq c$ were introduced in F at position T^i , and likewise for the b_i in $\phi(F)$. Using Lemma D.10, as $a \simeq_{n-1} b$, we conclude that $a_i \simeq_n b_i$, hence $a_i \leq_{n-1} b_i$.

This concludes the proof.

E. Proofs for Section VI: Arbitrary UIDs: Lifting Assumption Reversible

This appendix proves the claims needed to complete our proof of Theorem III.6, the existence of universal instances for UIDs, UFDs, and acyclic CQs of fixed size. The main claim is the existence of manageable partitions (Lemma VI.5).

Remember that we are assuming the “Unique Witness Property” (Section II) and that the constraints Σ_{U} are closed under the finite closure rule (in particular, Σ_{UID} is transitively closed).

E.1. Finite closure computation algorithm

For convenience we recall here how the finite closure is computed, from [8].

Given a set $\Sigma = \Sigma_{\text{FD}} \sqcup \Sigma_{\text{UID}}$ of FDs and UIDs, an ID *path* of Σ is a sequence of UIDs of Σ_{UID} of the following form: $R_1^{i_1} \subseteq R_2^{j_2}, R_2^{i_2} \subseteq R_3^{j_3}, \dots, R_{n-1}^{i_{n-1}} \subseteq R_n^{j_n}$, with $i_k \neq j_k$ for all k . The path is *functional* if, for all $1 < k < n$, $R_k^{i_k} \rightarrow R_k^{j_k} \in \Sigma_{\text{FD}}$. Note that our definition of the \rightarrow relation ensures that $\tau \rightarrow \tau'$ iff τ, τ' is a functional ID path.

An *invertible cycle* C of Σ is a functional ID path with $R_n = R_1$ and $j_n = j_1$ (so that $R_1^{i_1} \rightarrow R_1^{j_1} \in \Sigma_{\text{FD}}$): a UID that occurs in an invertible cycle is said to be *invertible*. The *reverse* \overline{C} of an invertible cycle C is $R_n^{j_n} \subseteq R_{n-1}^{i_{n-1}}, \dots, R_2^{j_2} \subseteq R_1^{i_1}$.

Applying the *cycle closure rule* in Σ means taking every invertible cycle C of Σ and adding to Σ the UIDs and UFDs needed to make \overline{C} an invertible cycle in Σ , namely, $R_2^{j_2} \subseteq R_1^{i_1}, R_3^{j_3} \subseteq R_2^{i_2}, \dots, R_n^{j_n} \subseteq R_{n-1}^{i_{n-1}}$, and $R_k^{j_k} \rightarrow R_k^{i_k}$ for $1 \leq k \leq n$. The finite closure is computed by closing under the rule above and by implication of the UIDs and of the FDs in isolation.

The fact that the result is exactly the finite closure of Σ is shown in [8].

E.2. Proof of Lemma VI.2 (New violations follow \rightarrow)

Lemma VI.2. *Let J be an aligned superinstance of I_0 and J' be the result of applying a thrifty chase step on J for a UID τ of Σ_{UID} . Assume that a UID τ' of Σ_{UID} was satisfied by J but is not satisfied by J' . Then $\tau \rightarrow \tau'$.*

Fix J, J' and $\tau : R^p \subseteq S^q$ and τ' . As chase steps add a single fact, the only new UID violations in J' relative to I are on elements in the newly created fact $F_n = S(\mathbf{b})$. As Σ_{UID} is transitively closed, F_n can introduce no new violation on the exported element b_q . Now, as thrifty chase steps always reuse existing elements at non-dangerous positions, we know that if $S^r \in \text{NDng}(S^q)$ then no new UID can be applicable to b_r . Hence, if a new UID is applicable to b_r for $S^r \in \text{Pos}(S)$, then necessarily $S^r \in \text{Dng}(S^q)$. By definition of dangerous positions, the UFD $S^r \rightarrow S^q$ is in Σ_{UFD} , and it is non-trivial because $S^r \neq S^q$. Hence, writing $\tau' : S^r \subseteq T^r$, we see that $\tau \rightarrow \tau'$.

E.3. Proof of Corollary VI.4 (Dealing with trivial classes)

Corollary VI.4. *For any trivial class $\{\tau\}$, performing one chase round on an aligned fact-saturated superinstance J of I_0 by fresh fact-thrifty chase steps for τ yields an aligned superinstance J' of I_0 that satisfies τ .*

Fix J, J' and τ . All violations of τ in J have been satisfied in J' by definition of J' , so we only have to show that no new violations of τ were introduced in J' . But by Lemma VI.2, as $\tau \not\rightarrow \tau$, each fresh fact-thrifty chase step cannot introduce such a violation, hence there is no new violation of τ in J' . Hence, $J' \models \tau$.

E.4. Proof of Lemma VI.5 (Existence of manageable partitions)

Our goal in this section is to show:

Lemma VI.5. *Any conjunction Σ_{UID} of UIDs closed under finite implication has a manageable partition.*

We assume that Σ_{UID} is closed under the finite closure rule (see Appendix E.1). Hence, in particular, it is transitively closed.

We start by introducing definitions about the \rightarrow relation, which we recall is defined so that $\tau \rightarrow \tau'$ for $\tau, \tau' \in \Sigma_{\text{UID}}$ whenever τ, τ' is a functional ID path, namely: letting $\tau : R^p \subseteq S^q$ and $\tau' : S^r \subseteq T^u$, the UFD $S^r \rightarrow S^q$ is non-trivial and is in Σ_{UFD} .

We extend \rightarrow to sets of UIDs in the expected way: $P \rightarrow P'$ if there exists $\tau \in P, \tau' \in P'$ such that $\tau \rightarrow \tau'$.

Definition E.1. The ID **graph** $\Gamma(\Sigma_{\text{UID}})$ is the directed graph (with self-loops) defined on Σ_{UID} by the \rightarrow relation. We define the **strongly connected components** of $\Gamma(\Sigma_{\text{UID}})$ as usual: an SCC is a maximal subset P of Σ_{UID} such that for all $\tau, \tau' \in P$, we have $\tau \rightarrow^* \tau'$, where \rightarrow^* denotes the transitive and reflexive closure of the \rightarrow relation. The **SCC graph** $G(\Sigma_{\text{UID}})$ is the directed acyclic graph (without self-loops) defined on the SCCs of $\Gamma(\Sigma_{\text{UID}})$ such that, for any two SCCs $P \neq P'$ of $\Gamma(\Sigma_{\text{UID}})$, there is an edge from P to P' iff $P \rightarrow P'$.

Note that the definition of SCCs allows both singleton SCCs $\{\tau\}$ where we have a self-loop ($\tau \rightarrow \tau$), and singletons where there is none ($\tau \not\rightarrow \tau$). We say that an SCC is **trivial** if it is a singleton without self-loops. Otherwise, if the SCC is not a singleton or if it has a self-loop, we call it **non-trivial**.

We first show the following lemma to understand the structure of the SCCs of $\Gamma(\Sigma_{\text{UID}})$. This lemma is proved in Appendix E.5.

Lemma E.2 (SCC structure). *The SCCs of $\Gamma(\Sigma_{\text{UID}})$ are transitively closed sets of UIDs. Further, for any non-trivial SCC P , letting $P^{-1} := \{\tau^{-1} \mid \tau \in P\}$, all UIDs of P^{-1} are in Σ_{UID} , and P^{-1} is an SCC of $\Gamma(\Sigma_{\text{UID}})$.*

Note that P and P^{-1} , as SCCs of $\Gamma(\Sigma_{\text{UID}})$, may be equal or disjoint. We accordingly call **self-inverse** an SCC P that is non-trivial but satisfies $P = P^{-1}$; non-trivial SCCs such that P and P^{-1} are disjoint are called **non-self-inverse**.

Given the structure of the SCCs, the first step to construct a manageable partition is to construct a topological sort of the SCC graph $G(\Sigma_{\text{UID}})$ of $\Gamma(\Sigma_{\text{UID}})$, but with an additional property, motivated by what we showed in Lemma E.2:

Definition E.3. A topological sort of $G(\Sigma_{\text{UID}})$ is **inverse-sequential** if, for any non-self-inverse SCC P , the SCCs P and P^{-1} are enumerated consecutively.

The first result, proven in Appendix E.6, is to justify that we can indeed construct an inverse-sequential topological sort of the SCC graph of $\Gamma(\Sigma_{\text{UID}})$:

Proposition E.4 (Inverse-sequential topological sort). *For any conjunction Σ_{UID} of UIDs closed under finite implication, $G(\Sigma_{\text{UID}})$ has an inverse-sequential topological sort.*

The second step is to construct the manageable partition itself from the inverse-sequential topological sort. Here is how we define the ordered partition from the topological sort:

Definition E.5. An inverse-sequential topological sort defines an ordered partition (P_1, \dots, P_n) of Σ_{UID} , in the following way: each class P_i of the partition either corresponds to one SCC of $G(\Sigma_{\text{UID}})$ (which is either trivial or self-inverse), or to the union of an SCC and its inverse SCC (which were enumerated consecutively because the topological sort is inverse-sequential). It is immediate that (P_1, \dots, P_n) is indeed an ordered partition, as it is constructed from a topological sort by merging some classes that were enumerated consecutively.

The second result is to show that the resulting ordered partition is indeed a manageable partition. In other words, we must show that the classes of the partitions are either trivial, or that they are a set of UID that is transitively closed and satisfies assumption reversible.

Proposition E.6 (Manageable partitions from sorts). *For any conjunction Σ_{UID} of UIDs closed under finite implication, letting \mathbf{P} be an ordered partition obtained from an inverse-sequential topological sort of $G(\Sigma_{\text{UID}})$, \mathbf{P} is a manageable partition.*

This second result is proven in Appendix E.7 and concludes the proof of our original claim.

E.5. Proof of the SCC Structure Lemma (Lemma E.2)

Lemma E.2 (SCC structure). *The SCCs of $\Gamma(\Sigma_{\text{UID}})$ are transitively closed sets of UIDs. Further, for any non-trivial SCC P , letting $P^{-1} := \{\tau^{-1} \mid \tau \in P\}$, all UIDs of P^{-1} are in Σ_{UID} , and P^{-1} is an SCC of $\Gamma(\Sigma_{\text{UID}})$.*

We first show an general lemma:

Lemma E.7. *Let P be a non-trivial SCC of $\Gamma(\Sigma_{\text{UID}})$. For any $\tau, \tau' \in P$, there is an invertible cycle of UIDs of P in which τ and τ' occur.*

Proof. Because P is a non-trivial SCC, we have $\tau \rightarrow^* \tau'$ and $\tau' \rightarrow^* \tau$, and the desired invertible cycle is obtained by concatenating the functional ID paths from τ to τ' , and from τ' to τ . Because P is an SCC, it is immediate that the UIDs of the resulting path are all in P . \square

We then divide our claim in two lemmas:

Lemma E.8. *Let P be an SCC of $\Gamma(\Sigma_{\text{UID}})$. Then P is closed under the transitivity rule.*

Proof. Let P be an SCC. If P consists of a single UID, then transitivity is immediately respected, so we assume that P contains > 1 UIDs. In particular, P is non-trivial. Let $\tau : R^p \subseteq S^q$ and $\tau' : S^q \subseteq T^r$ be two UIDs of P with $R^p \neq T^r$. As Σ_{UID} is closed under transitivity, we know $\tau'' : R^p \subseteq T^r$ is in Σ_{UID} . We show that $\tau'' \in P$.

As P is a non-trivial SCC, there is a functional ID path $\tau' = \tau_1 \rightarrow \dots \rightarrow \tau_n = \tau$, where $\tau_i \in P$ for all $1 \leq i \leq n$. Because of the UFDs that must be in Σ_{UFD} to make it a functional ID path, it is immediate that the following two paths are functional ID paths as well: $\tau'' \rightarrow \tau_2 \rightarrow \dots \rightarrow \tau_n$ and $\tau_1 \rightarrow \dots \rightarrow \tau_{n-1} \rightarrow \tau''$. Thus we have $\tau'' \rightarrow^* \tau$, and $\tau' \rightarrow^* \tau''$ where $\tau, \tau' \in P$, so that $\tau'' \in P$ by definition of an SCC. \square

Lemma E.9. *Let P be a non-trivial SCC of $\Gamma(\Sigma_{\text{UID}})$, and let $P^{-1} := \{\tau^{-1} \mid \tau \in P\}$. Then $P^{-1} \subseteq \Sigma_{\text{UID}}$, and P^{-1} is an SCC of $\Gamma(\Sigma_{\text{UID}})$.*

Proof. We first prove that, for any $\tau \in P$, $\tau^{-1} \in \Sigma_{\text{UID}}$. This is a direct consequence of Lemma E.7: there is an invertible cycle of P containing τ , so that by definition of an invertible cycle, τ^{-1} is in Σ_{UID} . We now turn to the second part of the claim.

First, we show that for any two $\tau, \tau' \in P^{-1}$, there is a functional ID path from τ to τ' , so that P^{-1} is strongly connected. This is clear: by Lemma E.7, there exists an invertible cycle C of P containing τ^{-1} and $(\tau')^{-1} \in P$, and the reverse \overline{C} of this cycle is also an invertible cycle, because Σ_{U} is finitely closed; \overline{C} is then a cycle of UIDs of P^{-1} containing τ and τ' .

Second, we show that for any UID $\tau \in \Sigma_{\text{UID}}$, if $P^{-1} \rightarrow^* \tau$ and $\tau \rightarrow^* P^{-1}$ then $\tau \in P^{-1}$. Consider such a UID τ , and let $p_1 : \tau' = \tau'_1 \rightarrow \dots \rightarrow \tau'_n = \tau$ and $p_2 : \tau = \tau''_1 \rightarrow \dots \rightarrow \tau''_m = \tau''$ be the witnessing functional ID paths, with $\tau', \tau'' \in P^{-1}$. We showed in the previous paragraph that P^{-1} is strongly connected: consider a (possibly empty) functional ID path p_3 from τ'' to τ' witnessing the fact that $\tau'' \rightarrow^* \tau$. Concatenating p_1 , p_2 and p_3 yields an invertible cycle C , so that because Σ_{U} is finitely closed, its reverse \overline{C} is also an invertible cycle. But \overline{C} witnesses the fact that $(\tau'')^{-1} \rightarrow^* \tau^{-1}$ and $\tau^{-1} \rightarrow^* (\tau')^{-1}$. Now, as $(\tau')^{-1}, (\tau'')^{-1} \in P$ and P is an SCC, we have $\tau^{-1} \in P$, so that $\tau \in P^{-1}$, the desired claim. Hence, P^{-1} is both strongly connected and maximal, so it is an SCC. \square

This concludes the proof. Note that, as P and P^{-1} are both SCCs of $\Gamma(\Sigma_{\text{UID}})$, either they are equal or they are disjoint. We observe that both cases may occur:

Example E.10. Consider the UIDs $\tau : R^2 \subseteq S^2$ and $\tau' : S^1 \subseteq R^1$, and the UFDs $\phi : R^2 \rightarrow R^1$ and $\phi' : S^1 \rightarrow S^2$. τ, τ' is an invertible cycle, so that by the finite closure rule, the UIDs τ^{-1} and $(\tau')^{-1}$ and the reverse UFDs are implied. However in $\Gamma(\Sigma_{\text{UID}})$ we have $\tau \succ \tau'$, $\tau' \succ \tau$, $\tau^{-1} \succ (\tau')^{-1}$, $(\tau')^{-1} \succ \tau^{-1}$, so that $\{\tau, \tau'\}$ and $\{\tau^{-1}, (\tau')^{-1}\}$ are two disjoint SCCs.

Consider now the UIDs $\tau : R^2 \subseteq S^2$, $\tau^{-1} : S^2 \subseteq R^2$, $\tau' : R^1 \subseteq R^3$, $\tau'' : S^3 \subseteq S^1$, and the UFDs $R^1 \rightarrow R^2$, $R^2 \rightarrow R^3$, $S^3 \rightarrow S^2$ and $S^2 \rightarrow S^1$. We can construct the invertible cycles τ' and τ'' , so that $(\tau')^{-1}$ and $(\tau'')^{-1}$ are implied by the finite closure rule. However, besides $\tau' \succ \tau'$, $\tau'' \succ \tau''$, $(\tau')^{-1} \succ (\tau')^{-1}$, $(\tau'')^{-1} \succ (\tau'')^{-1}$, it is also the case that $\tau \succ \tau''$, $\tau'' \succ \tau^{-1}$, $\tau^{-1} \succ \tau'$ and $\tau' \succ \tau$, and using the reverse UFDs the same is true of the inverses of τ' , $(\tau')^{-1}$, τ'' , and $(\tau'')^{-1}$. So in fact there is only one SCC $P = \{\tau, \tau^{-1}, \tau', (\tau')^{-1}, \tau'', (\tau'')^{-1}\}$, with $P^{-1} = P$.

E.6. Proof of the Inverse-Sequential Topological Sort Proposition (Proposition E.4)

We now prove that $G(\Sigma_{\text{UID}})$ has an inverse-sequential topological sort:

Proposition E.4 (Inverse-sequential topological sort). *For any conjunction Σ_{UID} of UIDs closed under finite implication, $G(\Sigma_{\text{UID}})$ has an inverse-sequential topological sort.*

For this we need the following observation about $G(\Sigma_{\text{UID}})$:

Lemma E.11. *Let P be a non-self-inverse SCC and consider $\tau \in \Sigma_{\text{UID}} \setminus (P \cup P^{-1})$ such that $\tau \succ P$. Then one of the following holds:*

- we have $\tau \succ P^{-1}$
- the SCC of τ is trivial, and for any $\tau_p \in \Sigma_{\text{UID}}$ such that $\tau_p \succ \tau$, we have $\tau_p \succ^* P^{-1}$.

Proof. Fix $\tau \in \Sigma_{\text{UID}} \setminus (P \cup P^{-1})$ and assume that we have $\tau \succ P$, i.e., $\tau \succ \tau'$ for some $\tau' \in P$. As P is non-trivial, using Lemma E.7, consider the predecessor τ'_{n-1} of τ' in an invertible cycle containing τ' (possibly $\tau'_{n-1} = \tau'$). Let R^p be the second position of τ , R^q be the first position of τ' , and R^r be the second position of τ'_{n-1} . Note that we have $R^r \neq R^q$ because $\tau'_{n-1} \succ \tau'$, and $R^p \neq R^q$ because $\tau \succ \tau'$. Observe that if $R^p \neq R^r$, then $\tau \succ (\tau'_{n-1})^{-1}$ because $R^r \rightarrow R^q$ and $R^q \rightarrow R^p$ hold in Σ_{UFD} (as these UFDs are used in an invertible cycle) and Σ_{UFD} is closed under transitivity. This proves the claim, as taking $\tau'' := (\tau'_{n-1})^{-1} \in P^{-1}$, we have $\tau \succ \tau''$.

If $R^p = R^r$, let P' be the SCC of τ . Assume first that P' is non-trivial. In this case, by Lemma E.7, there is an invertible cycle $\tau = \tau_1, \dots, \tau_m = \tau$ in P' . But then, we have $\tau'_{n-1} \succ \tau_2$, so that $P \succ P'$, and as $P' \succ P$ we have $P = P'$, so $\tau \in P$, a contradiction.

Hence, P' is trivial. Let S^q be the first position of τ and T^u be the first position of τ'_{n-1} . We must have $S^q \neq T^u$, as otherwise we have $\tau = \tau'_{n-1}$ so $\tau \in P$, a contradiction. Hence, because $(\tau'_{n-1})^{-1}$ is in Σ_{UID} (as $\tau'_{n-1} \in P$), by transitivity $\tau'' : S^q \subseteq T^u$ is in Σ_{UID} . We can then see that $\tau'' \succ (\tau'_{n-2})^{-1}$ because we had $(\tau'_{n-1})^{-1} \succ (\tau'_{n-2})^{-1}$ and both UIDs share the same second position; hence, $\tau'' \succ P^{-1}$. Now as τ'' and τ have the same first position, for any $\tau_p \in \Sigma_{\text{UID}}$, clearly $\tau_p \succ \tau$ implies that $\tau_p \succ \tau'' \succ P^{-1}$, proving the last part of the claim. \square

We now construct the inverse-sequential topological sort of $G(\Sigma_{\text{UID}})$ by enumerating the SCCs in a certain way that respects the \succ relation and maintains the following invariant: whenever P is non-self-inverse, then P and P^{-1} are enumerated consecutively; this guarantees that the result is a topological sort and that it is inverse-sequential.

First, whenever trivial or self-inverse SCCs can be enumerated, enumerate them. Second, whenever the SCCs that can be enumerated are all non-self-inverse, choose one such P to enumerate. By the invariant, P^{-1} has not yet been enumerated, otherwise P would have been enumerated immediately after. We want to enumerate P , and then enumerate P^{-1} .

To see why this is doable, we must show that, assuming that $P \neq P^{-1}$, if P can be enumerated and no trivial or self-inverse SCCs can be enumerated, then P^{-1} can also be enumerated. Let P' be a parent SCC of P^{-1} in $G(\Sigma_{\text{UID}})$ (so that $P' \succrightarrow P^{-1}$), and show that it has been enumerated already. If we have $P' = P$, meaning that $P \succrightarrow P^{-1}$, then this is not a problem, because we are about to enumerate first P and then P^{-1} , so we may assume that $P' \neq P$. Hence, P' is different from P and P^{-1} , so it is disjoint from it. We apply Lemma E.11 to any $\tau \in P'$. In the first case, we also have $P' \succrightarrow P$, so as P can be enumerated, P' was enumerated already. In the second case, $P' = \{\tau\}$ is trivial; further, considering any $P'' \succrightarrow P'$, we have $P'' \succrightarrow^* P$, so P'' was enumerated already. Hence, all such P'' are already enumerated, so that P' can be enumerated, but as it is trivial, it must have been enumerated already. Hence, in both cases P' was already enumerated unless it is P . This ensures that we can indeed enumerate P and P^{-1} consecutively, maintaining our invariant. Thus, we have constructed an inverse-sequential topological sort of $G(\Sigma_{\text{UID}})$. This concludes the proof.

E.7. Proof of the Manageable Partitions From Sorts Proposition (Proposition E.6)

Proposition E.6 (Manageable partitions from sorts). *For any conjunction Σ_{UID} of UIDs closed under finite implication, letting \mathbf{P} be an ordered partition obtained from an inverse-sequential topological sort of $G(\Sigma_{\text{UID}})$, \mathbf{P} is a manageable partition.*

Let (P_1, \dots, P_n) be the ordered partition. We prove that it is manageable. Trivial SCCs are indeed trivial classes of the partition, so we must only justify that any other class P_i is transitively closed and satisfies assumption reversible.

We define $\text{Pos}(P)$ for P a set of UIDs as the set of positions occurring in P , as in the definition of assumption reversible. We first prove a general lemma to take care of the second part of the assumption:

Lemma E.12. *Let P be a non-trivial SCC of $\Gamma(\Sigma_{\text{UID}})$. For any two positions $R^i \neq R^j$ of $\text{Pos}(P)$, if $R^i \rightarrow R^j$ is in Σ_{UFD} then so is $R^j \rightarrow R^i$.*

Proof. Fix R^i and R^j , assume that $\phi : R^i \rightarrow R^j$ is in Σ_{UFD} , and show that $\phi^{-1} : R^j \rightarrow R^i$ also is. Let τ_i be a UID of P where R^i occurs, and τ_j be a UID of P where R^j occurs. By Lemma E.7, there exists an invertible cycle C_1 where R^i and R^j occur.

We write $C_1 = R_1^{i_1} \subseteq R_2^{j_2}, \dots, R_n^{i_n} \subseteq R_1^{j_1}$, with some $1 \leq p, q \leq n$ such that $R_p = R_q = R$, and either $i_p = i$ or $j_p = i$, and either $i_q = j$ or $j_q = j$. By definition of an invertible cycle, the UFDs $\phi_p : R^{i_p} \rightarrow R^{j_p}$, $\phi_p^{-1} : R^{j_p} \rightarrow R^{i_p}$, $\phi_q : R^{i_q} \rightarrow R^{j_q}$ and $\phi_q^{-1} : R^{j_q} \rightarrow R^{i_q}$ are in Σ_{UFD} . Thus, because Σ_{UFD} is closed under transitivity, it is clear that if two positions among $S = (R^{j_p}, R^{i_p}, R^{j_q}, R^{i_q})$ are equal (in particular, if $p = q$), then we have $R^x \leftrightarrow_{\text{FUN}} R^y$ for any two positions R^x, R^y in S . Hence, as we know that $R^i \neq R^j$, and R^i and R^j are in S , the only case where we cannot conclude is the one where all the positions of S are different.

If all positions of S are different, then, because of $\phi_p, \phi_q, \phi_p^{-1}$ and ϕ_q^{-1} , by transitivity of Σ_{UFD} , we know that for any $x_1, x_2 \in \{i_p, j_p\}, y_1, y_2 \in \{i_q, j_q\}$, the UFD $R^{x_1} \rightarrow R^{y_1}$ is in Σ_{UFD} iff the UFD $R^{x_2} \rightarrow R^{y_2}$ is. Hence, since ϕ is in Σ_{UFD} , as $i \in \{i_p, j_p\}$ and $j \in \{i_q, j_q\}$, we know that $R^x \rightarrow R^y$ is in Σ_{UFD} for all $x \in \{i_p, j_p\}, y \in \{i_q, j_q\}$, and, to prove that ϕ^{-1} is in Σ_{UFD} , it suffices to show that $R^y \rightarrow R^x$ is in Σ_{UFD} for some $x \in \{i_p, j_p\}, y \in \{i_q, j_q\}$.

So let us construct the cycle $C_2 = R_1^{i_1} \subseteq R_2^{j_2}, \dots, R_{q-1}^{i_{q-1}} \subseteq R_q^{j_q}, R_p^{i_p} \subseteq R_{p+1}^{j_{p+1}}, \dots, R_n^{i_n} \subseteq R_1^{j_1}$. This is an invertible cycle, because $R_q = R_p = R$, and $R^{i_p} \neq R^{j_q}$ and the FD $R^{i_p} \rightarrow R^{j_q}$ is in Σ_{UFD} by our assumption. Hence, as C_2 is an invertible cycle, and because Σ_U is finitely closed, the reverse FD $R^{j_q} \rightarrow R^{i_p}$ is in Σ_{UFD} , which implies that ϕ^{-1} is in Σ_{UFD} . \square

We then show a lemma to help justify that the classes are transitively closed:

Lemma E.13. *For any non-trivial SCC P , if there is $\tau \in P$ and $\tau' \in P^{-1}$ such that $\tau^{-1} \neq \tau'$ but the second position of τ is the first position of τ' , then $P = P^{-1}$.*

Proof. We first observe that we have $P \rightarrow^* P^{-1}$. Indeed, as P and P^{-1} are non-trivial, consider $\tau_0 \in P$ and $\tau'_0 \in P^{-1}$ such that $\tau_0 \rightarrow \tau$ and $\tau' \rightarrow \tau'_0$. Letting τ'' be the UID which is transitively implied by τ and τ' , we know that it must be in Σ_{UID} as it is transitively closed, and we observe that $\tau_0 \rightarrow \tau'' \rightarrow \tau'_0$, so that $P \rightarrow^* P^{-1}$.

Now, write $\tau : R^p \subseteq S^q$ and $\tau' : S^q \subseteq T^r$, with $R^p \neq T^r$. As P and P^{-1} are non-trivial, using Lemma E.7, we can consider a functional ID path $\tau = \tau_1 \rightarrow \tau_2 \rightarrow \dots \rightarrow \tau_n = (\tau')^{-1}$, and a functional ID path $\tau^{-1} = \tau'_1 \rightarrow \dots \rightarrow \tau'_m = \tau'$. By Lemma E.12, all UFDs along these paths are such that their reverses are also in Σ_{UFD} . Consider now the smallest $k \geq 2$ such that we have $\tau_k^{-1} \neq \tau'_{m-k+1}$; such a k must exist because we have $\tau_n = (\tau')^{-1}$ and $\tau'_1 = \tau^{-1}$, and we know that $\tau^{-1} \neq \tau'$. Consider $\tau'' := \tau_k \in P$, and $\tau''' := \tau'_{m-k+1} \in P^{-1}$, and let S^u and S^v be respectively the first position of τ'' and the second position of τ''' : indeed it is easily observed that these positions must be in the same relation S , as this is true for τ_2 and τ'_{m-1} and is preserved for τ'' and τ''' because we have $\tau_l^{-1} = \tau'_{m-l+1}$ for all $1 \leq 2 \leq k$.

We now distinguish two cases. The first case is $S^v \neq S^u$, and we then have $\tau''' \rightarrow \tau''$, so that $P^{-1} \rightarrow P$. The second case is $S^v = S^u$. In this case, τ''' and τ'' are two UIDs of P^{-1} and P such that $(\tau''')^{-1} \neq \tau''$ but the second position of τ''' is the first position of τ'' . Hence, applying the reasoning of the first paragraph to τ'' and τ , we deduce that $P^{-1} \rightarrow^* P$. In either case, as we observed initially that $P \rightarrow^* P^{-1}$, we conclude that $P = P^{-1}$, the desired claim. \square

Corollary E.14. *For any non-trivial SCC P , $P \cup P^{-1}$ is transitively closed.*

Proof. By Lemma E.8, P and P^{-1} are transitively closed. Hence, if no UIDs is transitively implied by one UID from P and one from P^{-1} (or one from P^{-1} and one from P), then the claim is proven. Otherwise, by Lemma E.13, we have $P = P^{-1}$, so we can conclude by applying Lemma E.8 to $P = P \cup P^{-1}$. \square

We now conclude the proof of Proposition E.6. Let P_i be a class of the ordered partition (P_1, \dots, P_n) . We must show that it is either trivial or reversible. If it is not trivial, then we must show three things:

- P_i is transitively closed
- For every $\tau \in P_i$, we have $\tau^{-1} \in P_i$.
- For every two positions $R^p, R^q \in \text{Pos}(P_i)$ such that $R^p \rightarrow R^q$ is in Σ_{UFD} , $R^q \rightarrow R^p$ is also in Σ_{UFD} .

For the first claim, as P_i is not trivial, it is either a self-inverse SCC P of $\Gamma(\Sigma_{\text{UID}})$ (and the claim follows by Lemma E.8) or it is a union $P \cup P^{-1}$ where P is a non-self-inverse SCC (and the claim follows by Corollary E.14). The second claim is immediate by construction. The third claim is what is shown by Lemma E.12, noting that for any SCC P of Σ_{UID} , we have $\text{Pos}(P) = \text{Pos}(P^{-1})$. This concludes the proof of Proposition E.6.

F. Proofs for Section VII: Higher-Arity FDs

In this section, we show what is needed to adapt the Acyclic Unary Universal Models Theorem (Theorem III.6) to produce aligned superinstances that satisfy the full set of constraints Σ rather than just the unary subset Σ_U .

F.1. Proof of the Sufficiently Envelope-Saturated Solutions Proposition (Proposition VII.5)

We now prove the following result, which provides our way to construct the initial instance on which we apply the completion process of the previous sections:

Proposition VII.5 (Sufficiently envelope-saturated solutions). *For any $K \in \mathbb{N}$ and instance I_0 , we can build a superinstance I'_0 of I_0 that is k -sound for CQ, and an aligned superinstance J of I'_0 that satisfies Σ_{FD} and is $(K|J|)$ -envelope-saturated.*

We define the notation $|\sigma| := \max_{R \in \sigma} |R|$, and also define the following:

Definition F.1. The **overlap** $\text{OVL}(F, F')$ between two facts $F = R(\mathbf{a})$ and $F' = R(\mathbf{b})$ of the same relation R in an instance I is the subset O of $\text{Pos}(R)$ such that $a_s = b_s$ iff $R^s \in O$. If $|O| > 0$, we say that F and F' **overlap**.

We also define the following, which are the FDs used in the definition of envelopes (Definition VII.2):

Definition F.2. Given a set Σ_{FD} of FDs on a relation R and $O \subseteq \text{Pos}(R)$, the FD **projection** Σ_{FD}^O of Σ_{FD} to O are the FDs $R^L \rightarrow R^r$ of Σ_{FD} such that $R^L \subseteq O$ and $R^r \in O$, plus, for every FD $R^L \rightarrow R^r$ of Σ_{FD} where $R^L \subseteq O$ and $R^r \notin O$, the key dependency $R^L \rightarrow O$.

We first note the following immediate consequence of the Dense Interpretations Theorem (Theorem VII.7):

Corollary F.3. *We can assume in the Dense Interpretations Theorem (Theorem VII.7) that the resulting instance I is such that each element occurs at exactly one position of the relation R : formally, for all $a \in \text{dom}(I)$, there exists exactly one $R^p \in \text{Pos}(R)$ such that $a \in \pi_{R^p}(I)$.*

Proof. Create from I the instance I' whose domain is $\{(a, R^p) \mid a \in \text{dom}(I), R^p \in \text{Pos}(\sigma)\}$ and which contains for every fact $F = R(\mathbf{a})$ of I a fact $F' = R(\mathbf{b})$ such that $b_p = (a_p, R^p)$ for every $R^p \in \text{Pos}(\sigma)$. Clearly this defines a bijection ϕ from the facts of I to the facts of I' , and for any facts F, F' of I' , $\text{OVL}(F, F') = \text{OVL}(\phi^{-1}(F), \phi^{-1}(F'))$. Thus any violation of the FDs Σ_{FD} in I' would witness one in I . Of course, $|\text{dom}(I')| = |\sigma| \cdot |\text{dom}(I)|$, so that, letting K' be our target constant factor between $|\text{dom}(I')|$ and $|I|$, we must use $K := K'|\sigma|$ as the constant for the Dense Interpretation Theorem, so that $|I| \geq K'|\sigma| \cdot |\text{dom}(I)|$, which implies $|I'| \geq K'|\text{dom}(I')|$. \square

We also show two easy lemmas:

Lemma F.4. *Let I be an instance, Σ_{FD} be a conjunction of FDs, and $F \neq F'$ be two facts of I . Assume there is a position $R^p \in \text{Pos}(\sigma)$ such that, writing $O := \text{NDng}(R^p)$, we have $\text{OVL}(F, F') \subsetneq O$, and that $\{\pi_O(F), \pi_O(F')\}$ is not a violation of Σ_{FD}^O . Then $\{F, F'\}$ is not a violation of Σ_{FD} .*

Proof. Assume by way of contradiction that F and F' violate an FD $\phi : R^L \rightarrow R^r$ of Σ_{FD} , which implies that $R^L \subseteq \text{OVL}(F, F') \subseteq O$ and $R^r \notin \text{OVL}(F, F')$. Now, if $R^r \in O$, then ϕ is in Σ_{FD}^O , so that $\pi_O(F)$ and $\pi_O(F')$ violate Σ_{FD}^O , a contradiction. Hence, $R^r \in \text{Pos}(R) \setminus O$, and the key dependency $\kappa : R^L \rightarrow O$ is in Σ_{FD}^O , so that $\pi_O(F)$ and $\pi_O(F')$ must satisfy κ . Thus, because $R^L \subseteq \text{OVL}(F, F')$, we must have $\text{OVL}(F, F') = O$, which is a contradiction because we assumed $\text{OVL}(F, F') \subsetneq O$. \square

Lemma F.5. *For any $(R^p, C) \in \text{AFactCl}$, letting $O := \text{NDng}(R^p)$, if (R^p, C) is unsafe, then there is no position $R^q \in O$ that determines O in Σ_{FD}^O : formally, there is no $R^q \in O$ such that we have $R^q \rightarrow R^r$ in Σ_{FD}^O for all $R^r \in O$.*

Proof. Fix $D = (R^p, C)$ in AFactCl and let O be the non-dangerous positions of R^p . We first show that if Σ_{FD} implies that O has a unary key $R^s \in O$ in Σ_{FD} , then D is safe. Indeed, assume the existence of such a unary key R^s . If there were a FD $R^L \rightarrow R^r$ in Σ_{FD} with $R^L \subseteq O$ and $R^r \notin O$, then, by transitivity, the UFD $R^s \rightarrow R^r$ would be in Σ_{UFD} , which by Lemma C.2 implies that R^r is non-dangerous for R^p because $R^s \in O$ is non-dangerous for R^p . This contradicts our assumption that $R^r \notin O$.

We must now show that if O has a unary key in O according to Σ_{FD}^O then O has a unary key in O according to Σ_{FD} . It suffices to show that for any two positions $R^q, R^s \in O$, if $\phi : R^q \rightarrow R^s$ holds in Σ_{FD}^O then it also does in Σ_{FD} . Assuming to the contrary that there is such a ϕ , consider its derivation from the dependencies of Σ_{FD}^O . Clearly the derivation must be using one of the key dependencies $\kappa : R^L \rightarrow O$, which are the only dependencies in Σ_{FD}^O that are not in Σ_{FD} . But this means that, the first time we used such a dependency, we had derived a unary key dependency $R^q \rightarrow R^L$ using only the FDs of Σ_{FD} . Considering that κ was created to stand for a FD $R^L \rightarrow R^r$ in Σ_{FD} , with $R^r \notin O$, we deduce that we can derive from Σ_{FD} that $R^q \rightarrow R^r$, contradicting again the fact that $R^r \notin O$ (because R^r should then be in $\text{NDng}(R^p)$). Hence, if O has a unary key in O according to Σ_{FD}^O then D is safe. Thus, we have proven the contrapositive of the desired result. \square

We now prove Proposition VII.5. The bulk of the work is to show the following claim, for each unsafe class of AFactCl . The construction of global envelopes from the individual envelopes is then easy.

Lemma F.6. *For any unsafe class D in AFactCl and constant K , one can construct a superinstance I'_0 of I_0 that is k -sound for CQ, and an aligned superinstance $J = (I, \text{sim})$ of I'_0 that satisfies Σ_{FD} with an envelope E for D of size $K|J|$.*

Proof. Fix the unsafe achieved fact class $D = (R^p, C)$ and choose $F = R(\mathbf{b})$ a fact of $\text{Chase}(I_0, \Sigma_{\text{UID}}) \setminus I_0$ that achieves D . Let I_1 be obtained from I_0 by applying UID chase steps on I_0 to obtain a finite truncation of $\text{Chase}(I_0, \Sigma_{\text{UID}})$ that includes F but no child fact of F , and consider the aligned superinstance $J_1 = (I_1, \text{sim}_1)$ where sim_1 is the identity.

Let $O := \text{NDng}(R^p)$, and define a $|O|$ -ary relation $R_{|O|}$; for convenience, we index its positions by O . Because D is unsafe, by Lemma F.5, $R_{|O|}$ has no unary key in Σ_{FD}^O . Apply the Dense Interpretations Theorem (Theorem VII.7) to $R_{|O|}$ and Σ_{FD}^O with the additional condition of Corollary F.3, taking $K|J_1|$ as the constant. We thus obtain an instance I_D of $R_{|O|}$ that satisfies Σ_{FD}^O and such that, letting $N := |\text{dom}(I_D)|$, we have $|I_D| \geq NK|J_1|$. Let $I'_D \subseteq I_D$ be a subinstance of size N of I_D such that $\text{dom}(I'_D) = \text{dom}(I_D)$, that is, each element of $\text{dom}(I_D)$ occurs in some fact of I'_D . This can clearly be ensured by picking, for any element of $\text{dom}(I_D)$, one fact of I_D where it occurs, removing duplicate facts, and completing with other arbitrary facts of I_D to have N distinct facts. Number the facts of I'_D as F'_1, \dots, F'_N .

We create $N - 1$ disjoint copies of J_1 , numbered J_2 to J_N . We call $J' = (I', \text{sim})$ the disjoint union $J_1 \sqcup \dots \sqcup J_N$. It is clear that J' is indeed an aligned superinstance of I'_0 , where I'_0 is formed of the N disjoint copies of I_0 , and I'_0 is clearly a k -sound superinstance of I_0 for CQ. For $1 \leq i \leq N$, we call

$F_i = R(a^i)$ the fact of I_i that corresponds to the achiever F in $\text{Chase}(I_0, \Sigma_{\text{UID}})$. In particular, for all $1 \leq i \leq N$, we have that $\text{sim}(a_j^i) = b_j$ for all j , and a_p^i is the only element of F_i that also occurs in other facts of J_i .

We consider the application f that maps a_j^i , for $1 \leq i \leq N$ and $R^j \in O$, to $\pi_{R^j}(F_i')$. This application f is well-defined, because the a_j^i are pairwise distinct. We extend f to $\text{dom}(I')$, and call the extension f' , by setting $f'(a) := a$ if a is not in the domain of f . We call I the image of I' under f' . In other words, I is the underlying instance of J' except that elements at positions of O in the facts F_i were identified so that the projections to O of the $f'(F_i)$ are isomorphic to the F_i' . Because a_j^i occurs only in F_i for all $R^j \neq R^p$, and $R^p \notin O$, this means that the identified elements only occurred in the F_i in I' .

We now build $J = (I, \text{sim})$ obtained by defining sim from the sim_i as follows: any element a not in the domain of f is mapped to $\text{sim}_i(a)$ for the one i such that $a \in \text{dom}(I_i)$, and any a in the domain of f is mapped to $\text{sim}_i(a')$ for any preimage of a' by f . All that remains to show is that J is indeed an aligned superinstance of I'_0 satisfying the required conditions.

We note that it is immediate that J is a superinstance of I'_0 , as the achiever F is not a fact of I_0 , so that $\text{dom}(I_0)$ is not in the domain of f . It is clear that J has $N|J_1|$ facts, because, as $R^p \notin O$, no facts can be identified by f' . We now claim that J is an aligned superinstance of I'_0 , and that E , defined as the set of the tuples of I_D , is an envelope for I' and D . The fact that $|E| = K|J|$ is immediate.

The fact that sim is a k -bounded simulation from J to $\text{Chase}(I'_0, \Sigma_{\text{UID}})$ is by induction. The case of $k = 0$ is trivial. The induction case is trivial for all facts except for the $h'(F_i)$, because the a_j^i only occurred in I in the facts F_i , by our assumption that the F_i have no children in the I_i and by the fact that the exported position of F is $R^p \notin O$. Consider now one fact $F' = R(c)$ of I' which is the image by f' of a F_i . Choose $1 \leq p \leq |R|$. We show that there exists a fact $F'' = R(d)$ of $\text{Chase}(I'_0, \Sigma_{\text{UID}})$ such that $\text{sim}(c_p) = d_p$ and for all $1 \leq q \leq |R|$ we have $(I, c_q) \leq_{k-1} (\text{Chase}(I'_0, \Sigma_{\text{UID}}), d_q)$, which by induction hypothesis is implied by $\text{sim}(c_q) \simeq_k d_q$. Let $a_{j_0}^{i_0}$ be the preimage of a_p used to define $\text{sim}(a_p)$; by the condition of Corollary F.3, we must have $j_0 = p$. Consider the fact $F'' = R(d)$ of $\text{Chase}(I'_0, \Sigma_{\text{UID}})$ corresponding to F_{i_0} in I . By definition, $\text{sim}(c_p) = \text{sim}(a_{j_0}^{i_0}) = d_p$. Fix now $1 \leq q \leq |R|$. Let $a_{j'_0}^{i'_0}$ used to define $\text{sim}(c_q)$; again $j'_0 = q$ and $\text{sim}(c_q)$ is $\pi_{R^q}(F''')$ for the fact $F''' = R(e)$ of $\text{Chase}(I'_0, \Sigma_{\text{UID}})$ corresponding to $F_{i'_0}$ in I . But as both F''' and F'' are copies of the same achiever fact F of $\text{Chase}(I_0, \Sigma_{\text{UID}})$, we have $d_q \simeq_k e_q$, so that $\text{sim}(c_q) \simeq_k d_q$, what we wanted to show. This proves that sim is indeed a k -bounded simulation from J to $\text{Chase}(I_0, \Sigma_{\text{UID}})$.

We show that J satisfies Σ_{FD} . As I satisfies Σ_{FD} , any new violation of Σ_{FD} in I' relative to I must include some fact $F = h'(F'_{i_0})$, and some fact F' overlapping with F , so necessarily $F' = h'(F'_{i_1})$ for some i_1 by construction of I' , and $\text{OVL}(F, F') \subseteq O$. We now use Lemma F.4 to deduce that we cannot have $\text{OVL}(F, F') \subsetneq O$, so $\text{OVL}(F, F') = O$. By our definition of f and of the F_i' this implies that $F'_{i_0} = F'_{i_1}$, a contradiction because $F \neq F'$.

Thus, from the above, and as the technical conditions of the definition of aligned superinstances are clearly respected, J is indeed an aligned superinstance of I'_0 .

Last, we check that E is indeed an envelope. Indeed, it satisfies Σ_{FD}^O by construction, so the first two conditions are respected. The third condition is respected by the condition of Corollary F.3, and because the $f(a_j^i)$ always occur at position R^j in some fact of I'_D , as we constructed I'_D such that $\text{dom}(I'_D) = \text{dom}(I_D)$. The last condition is true because the envelope elements are only used in the $f(F_i)$, and the sim -images of the $f(F_i)$ are copies in $\text{Chase}(I'_0, \Sigma_{\text{UID}})$ of the same achiever fact F in $\text{Chase}(I_0, \Sigma_{\text{UID}})$.

Hence, J is indeed an aligned superinstance of a k -sound I'_0 that satisfies Σ_{FD} and has an envelope of size $K|J|$, proving the desired claim. \square

We now prove the main result by building I'_0 and the aligned superinstance $J = (I, \text{sim})$ of I'_0 that has a global envelope \mathcal{E} . As AFactCl is finite, we build one J_D per $D \in \text{AFactCl}$. When D is unsafe, we use the previous lemma. When $D = (R^p, C)$ is safe, we just take a single copy J_D of the truncated chase to achieve the class D , and take as the only fact of the envelope the projection to $\text{NDng}(R^p)$ of the fact of J_D corresponding to the achiever of D in $\text{Chase}(I_0, \Sigma_{\text{UID}})$. As AFactCl is finite and its size is a constant, we can ensure that $|\mathcal{E}(D)|$ for all unsafe $D \in \text{AFactCl}$ is $\geq (K+1)|I|$, by taking sufficiently large K when we apply Lemma F.6 for each unsafe class.

Let J be the disjoint union of the J_D . Each J_D is an aligned superinstance of an $(I'_0)_D$ which is a k -sound superinstance of I_0 . Hence, J is an aligned superinstance of the union of the $(I'_0)_D$ which is also k -sound. There are no violations of Σ_{FD} in J because there are none in any of the J_D , and the union is disjoint. The disjointness of domains of envelopes is because the J_D are disjoint. It is easy to see that J is $(K|I|)$ -envelope-saturated, because $|\mathcal{E}(D)| \geq (K+1)|I|$ for all unsafe $D \in \text{AFactCl}$, so the number of remaining facts of each envelope for an unsafe class is $\geq K|I|$ (every fact of I eliminates at most one fact in each envelope). Hence, the proposition is proven.

F.2. Proof of the Dense Interpretations Theorem (Theorem VII.7)

Remember that we want to show:

Theorem VII.7 (Dense interpretations). *For any set Σ_{FD} of FDs over a relation R with no unary key, and $K \in \mathbb{N}$, there exists a non-empty instance I of R that satisfies Σ_{FD} and has at least $K|\text{dom}(I)|$ facts.*

Fix the relation R , and let Σ_{FD} be an arbitrary set of FDs which we assume is closed under FD implication. Let Σ_{UFD} be the UFDs implied by Σ_{FD} ; it is also closed under FD implication. Recall the definition of OVL (Definition F.1). We introduce a notion of *safe overlaps* for Σ_{UFD} , which depends only on Σ_{UFD} but (we will show) is a sufficient condition to satisfy Σ_{FD} :

Definition F.7. *We say a subset $O \subseteq \text{Pos}(R)$ is **safe** for Σ_{UFD} if O is empty or for every $R^p \in \text{Pos}(R) \setminus O$, there exists $R^q \in \text{Pos}(R)$ such that the unary key dependency $R^q \rightarrow O$ is implied by Σ_{UFD} but the UFD $R^q \rightarrow R^p$ does not hold in Σ_{UFD} .*

*We say that an instance I has the **safe overlaps** property (for Σ_{UFD}) if for every $F \neq F'$ of I , $\text{OVL}(F, F')$ is safe.*

We now claim the following lemma, and its immediate corollary:

Lemma F.8. *If $O \subseteq \text{Pos}(R)$ is safe for Σ_{UFD} then there is no FD $\phi : R^L \rightarrow R^r$ in Σ_{FD} such that $R^L \subseteq O$ but $R^r \notin O$.*

Proof. If O is empty the claim is immediate. Otherwise, assume to the contrary the existence of such an FD ϕ . As $R^r \notin O$ and O is safe, there is $R^q \in \text{Pos}(R)$ such that $R^q \rightarrow O$ holds in Σ_{UFD} but $R^q \rightarrow R^r$ does not hold in Σ_{UFD} . Now, as $R^L \subseteq O$, we know that $R^q \rightarrow R^L$ holds in Σ_{UFD} , so that, by transitivity of Σ_{FD} , $\phi' : R^q \rightarrow R^r$ holds in Σ_{FD} . As ϕ' is a UFD, this implies it holds in Σ_{UFD} , a contradiction. \square

Corollary F.9. *For any instance I , if I has the safe overlaps property for Σ_{UFD} , then I satisfies Σ_{FD} .*

Proof. Considering two facts F and F' in I , as $\text{OVL}(F, F')$ is safe, we know that for any FD $\phi : R^L \rightarrow R^r$ in Σ_{FD} , we cannot have $R^L \subseteq O$ but $R^r \notin O$. Hence, F and F' cannot be a violation of ϕ . \square

Thus, it suffices to show the following generalization of the Dense Interpretations Theorem:

Theorem F.10. Let R be a relation and Σ_{UFD} be a set of UFDs over R . Let D be the number of positions of the smallest key of R for Σ_{UFD} : formally, $D := |K|$, where $K \subseteq \text{Pos}(R)$ is such that $R^K \rightarrow R^p$ holds in Σ_{UFD} for all $R^p \in \text{Pos}(R)$, and K has minimal cardinality among all subsets of $\text{Pos}(R)$ with this property. Let x be $\frac{D}{D-1}$ if $D > 1$ and 1 otherwise.

For every $N \geq 1$, there exists a finite instance I of R such that $|\text{dom}(I)|$ is $O(N)$, $|I|$ is $\Omega(N^x)$, and I has the safe overlaps property for Σ_{UFD} .

It is clear that this theorem implies the Dense Interpretations Theorem, because if R has no unary key for Σ_{FD} then $D > 1$ and thus $x > 1$, which implies that, for any K , by taking a sufficiently large N , we can obtain an instance I for R with N elements and KN facts that has the safe overlaps property for Σ_{UFD} ; now, by Lemma F.9, this implies that I satisfies Σ_{FD} .

We will now prove Theorem F.10. Fix the relation R and set of UFDs Σ_{UFD} . The case of $D = 1$ is vacuous and can be eliminated directly (consider the instance $\{R(a_i, \dots, a_i) \mid 1 \leq i \leq N\}$). Hence, assume that $D > 1$, and let $x := \frac{D}{D-1}$.

We first show the claim on a specific relation R_0 and set Σ_{UFD}^0 of UFDs. We will then generalize the construction to arbitrary relations and UFDs. Let $T_0 := \{1, \dots, D\}$, and consider a bijection $v : \{1, \dots, 2^D\} \rightarrow \mathfrak{P}(T_0) \setminus \{\emptyset\}$. Let R_0 be a $(2^D - 1)$ -ary relation, and take $\Sigma_{\text{UFD}}^0 := \{R^i \rightarrow R^j \mid v(i) \subseteq v(j)\}$. Note that Σ_{UFD}^0 is clearly closed under implication of UFDs. Fix $N \in \mathbb{N}$, and let us construct an instance I_0 with $O(N)$ elements and $\Omega(N^x)$ facts.

Fix $n := \lfloor N^{1/(D-1)} \rfloor$. Let \mathcal{F} be the set of partial functions from T_0 to $\{1, \dots, n\}$, and write $\mathcal{F} = \mathcal{F}_t \sqcup \mathcal{F}_p$, where \mathcal{F}_t and \mathcal{F}_p are respectively the total and the strictly partial functions. We take I_0 to consist of one fact F_f for each $f \in \mathcal{F}_t$, where $F_f = R_0(\mathbf{a}^f)$ is defined as follows: for $1 \leq i \leq 2^D$, $a_i^f := f|_{T_0 \setminus v(i)}$. In particular:

- $a_{v^{-1}(T_0)}^f$, the element of F_f at the position mapped to $T_0 \in \mathfrak{P}(T_0) \setminus \{\emptyset\}$, is the strictly partial function that is nowhere defined;
- $a_{\{i\}}^f$, the element of F_f at the position mapped to $\{i\} \in \mathfrak{P}(T_0) \setminus \{\emptyset\}$, is the strictly partial function equal to f except that it is undefined on i .

Hence, $\text{dom}(I_0) = \mathcal{F}_p$ (because \emptyset is not in the image of v), so that $|\text{dom}(I_0)| = \sum_{0 \leq i < D} \binom{D}{i} n^i$. Remembering that D is a constant, this implies that $|\text{dom}(I_0)|$ is $O(n^{D-1})$, so it is $O(N)$ by definition of n . Further, we claim that $|I_0| = |\mathcal{F}_t| = n^D = N^x$. To show this, consider two facts F_f and F_g , and show that $F_f = F_g$ implies $f = g$, so there are indeed $|\mathcal{F}_t|$ different facts in I_0 . As $\pi_{v^{-1}(\{1\})}(F_f) = \pi_{v^{-1}(\{1\})}(F_g)$, we have $f(t) = g(t)$ for all $t \in T_0 \setminus \{1\}$, and looking at $\pi_{v^{-1}(\{2\})}(F_f)$ and $\pi_{v^{-1}(\{2\})}(F_g)$ concludes (here we use the fact that $D \geq 2$). Hence, the cardinalities of I_0 and of its domain are suitable.

We must now show that I_0 has the safe overlaps property. For this we first make the following general observation:

Lemma F.11. Let Σ_{UFD} be any conjunction of UFDs and I be an instance such that $I \models \Sigma_{\text{UFD}}$. Assume that, for any pair of facts $F \neq F'$ of I that overlap, there exists $R^p \in \text{OVL}(F, F')$ which is a unary key for $\text{OVL}(F, F')$. Then I has the safe overlaps property for Σ_{UFD} .

Proof. Consider $F, F' \in I$ and $O := \text{OVL}(F, F')$. If $F = F'$, then $O = \text{Pos}(R)$, and O is clearly safe. Otherwise, if $F \neq F'$, let $R^p \in \text{Pos}(R) \setminus O$. Let $R^q \in O$ be the unary key of O . We know that $R^q \rightarrow O$ holds in Σ_{UFD} , so to show that O is safe it suffices to show that $\phi : R^q \rightarrow R^p$ does not hold in Σ_{UFD} . However, if it did, then as $R^q \in O$ and $R^p \notin O$, F and F' would witness a violation of ϕ , contradicting the fact that I satisfies Σ_{UFD} . \square

So we show that I_0 satisfies Σ_{UFD}^0 and that every non-empty overlap between facts of I_0 has a unary key.

First, to show that I_0 satisfies Σ_{UFD}^0 , observe that whenever $\phi : R_0^i \rightarrow R_0^j$ holds in Σ_{UFD} , then $v(i) \subseteq v(j)$, so that, for any fact F of I_0 , for any $1 \leq t \leq T_0$, whenever $(\pi_j(F))(t)$ is defined, so is $(\pi_i(F))(t)$, and we have $(\pi_j(F))(t) = (\pi_i(F))(t)$. Hence, letting F and F' be two facts of I_0 such that $\pi_i(F) = \pi_i(F')$, we know that $\pi_j(F)$ is defined iff $\pi_j(F')$ is (as this only depends on j), and, if both are defined, the previous observation shows that $\pi_j(F) = \pi_j(F')$. Hence, F and F' cannot witness a violation of ϕ .

Second, considering two facts $F_f = R_0(a^f)$ and $F_g = R_0(a^g)$, with $f \neq g$ so that $F_f \neq F_g$, we show that if $\text{OVL}(F_f, F_g)$ is non-empty then it has a unary key. Let $O := \{t \in T_0 \mid f(t) = g(t)\}$, and let $X = T_0 \setminus O$; we have $X \neq \emptyset$, because otherwise $f = g$, so we can define $p := v^{-1}(X)$. We will show that $\text{OVL}(F_f, F_g) = \{R^i \in \text{Pos}(R_0) \mid X \subseteq v(i)\}$. This implies that $R^p \in \text{OVL}(F_f, F_g)$ and that R^p is a unary key of $\text{OVL}(F_f, F_g)$, because, for all $R^q \in \text{OVL}(F_f, F_g)$, $X \subseteq v(R^q)$, so that $R^p \rightarrow R^q$ holds in Σ_{UFD} .

Indeed, consider R^i such that $X \subseteq v(i)$. Then $T_0 \setminus v(i) \subseteq T_0 \setminus X$, so that, because $a_i^f = f|_{T_0 \setminus v(i)}$ and $a_i^g = g|_{T_0 \setminus v(i)}$, we have $a_i^f = a_i^g$ by definition of $O = T_0 \setminus X$. Thus $R^i \in \text{OVL}(F_f, F_g)$. Conversely, if $R^i \in \text{OVL}(F_f, F_g)$, then we have $a_i^f = a_i^g$, so by definition of O we must have $T_0 \setminus v(i) \subseteq O' = T_0 \setminus X$, which implies $X \subseteq v(i)$.

Hence, I_0 is a finite instance of Σ_{UFD} which satisfies the safe overlaps property and contains $O(N)$ elements and $\Omega(N^{D/(D-1)})$ facts. This concludes the proof of Theorem F.10 for the specific case of R_0 and Σ_{UFD}^0 .

Let us now show the claim for the actual R and Σ_{UFD} . Let K be a key of R of minimal cardinality, so that $|K| = D$. Let λ be any bijective labeling from K to T_0 . Extend λ to a function μ from $\text{Pos}(R)$ to $\mathfrak{P}(T_0) \setminus \{\emptyset\}$ such that, for every $R^p \in \text{Pos}(R)$ and $R^k \in K$, we have $\lambda(R^k) \in \mu(R^p)$ iff $R^k = R^p$ or $R^k \rightarrow R^p$ holds in Σ_{UFD} .

Now, create the instance I of R from I_0 by creating, for every fact $F_0 = R_0(a)$ of I_0 , a fact $F = R(b)$ in I , with $b_i = a_{v^{-1}(\mu(R^i))}$ for all $1 \leq i \leq |R|$.

We do not create duplicate facts by the same argument as before, considering the projection of the facts of I to $R^{k_1} \neq R^{k_2}$ in K , because $\mu(R^{k_1}) = \{\lambda(R^{k_1})\}$ and $\mu(R^{k_2}) = \{\lambda(R^{k_2})\}$ (otherwise this contradicts the minimality of K). Hence I , as I_0 , has a suitable number of facts, and a suitable domain cardinality because $\text{dom}(I) \subseteq \text{dom}(I_0)$.

Let us now show that overlaps are safe in I . Consider two facts F, F' of I that overlap, and let $O := \text{OVL}(F, F')$. We first claim that there exists $\emptyset \subsetneq K' \subseteq K$, such that, letting $X' := \{\lambda(R^k) \mid R^k \in K'\}$, we have $\text{OVL}(F, F') = \{R^i \in \text{Pos}(R) \mid X' \subseteq \mu(R^i)\}$. Indeed, letting F_f and F_g be the facts of I_0 used to create F and F' , we previously showed the existence of $\emptyset \subsetneq X \subseteq T_0$ such that $\text{OVL}(F_f, F_g) = \{R^i \in \text{Pos}(R_0) \mid X \subseteq v(i)\}$. Our definition of F and F' from F_f and F_g makes it clear that we can satisfy the condition by taking $K' := \lambda^{-1}(X)$, so that $X' = X$.

Consider now $R^p \in \text{Pos}(R) \setminus O$. We cannot have $X' \subseteq \mu(R^p)$, otherwise $R^p \in O$. Hence, there exists $R^k \in K'$ such that $\lambda(R^k) \notin \mu(R^p)$. This implies that $R^k \rightarrow R^p$ does not hold in Σ_{UFD} . However, as $R^k \in K'$, we have $\lambda(R^k) \in \mu(R^q)$ for all $R^q \in O$, so that $R^k \rightarrow O$ holds in Σ_{UFD} . This proves that $O = \text{OVL}(F, F')$ is safe. Hence, I has the safe overlaps property, which concludes the proof.

F.3. Proof of Lemma VII.9 (Envelope-thrifty chase steps satisfy Σ_{FD})

Lemma VII.9. *For $n > 0$, for any n -envelope-saturated aligned superinstance J that satisfies Σ_{FD} , the result J' of an envelope-thrifty chase step on J is an $(n-1)$ -envelope-saturated superinstance that satisfies Σ_{FD} .*

Consider an application of an envelope-thrifty chase step: let $\tau : R^p \subseteq S^q$ be the UID, let $O := \text{NDng}(S^q)$, let $J = (I, \text{sim})$ be the aligned superinstance of I_0 , let $F_w = S(b')$ the chase witness, let

$D = (S^q, C)$ be the fact class, let $F_n = S(\mathbf{b})$ be the new fact to be created, and let \mathbf{t} be the remaining tuple of $\mathcal{E}(D)$ used to define F_n .

We first check that envelope-thrifty chase steps are well-defined in the sense that the fact class $D = (S^q, C)$ is indeed achieved in $\text{Chase}(I_0, \Sigma_{\text{UID}})$, so it is in AFactCl . To see why, observe that F_w is a fact of $\text{Chase}(I_0, \Sigma_{\text{UID}})$ whose fact class is (S^q, C) . Indeed, by Lemma D.2, b'_q is the exported element of F_w , and clearly $b'_i \in C_i$ for all $S^i \in \text{Pos}(S)$. Hence indeed $D \in \text{AFactCl}$.

It is then clear that envelope-thrifty chase steps are well-defined, in the sense that they are indeed thrifty chase steps: elements reused from the envelopes already occur at the positions where they are used in the new fact F_n . Further, their sim-image is the right one, by definition of an envelope.

We first prove that J' is still an aligned superinstance. This is shown exactly as in Lemma V.9, except for the fact that $J' \models \Sigma_{\text{UFD}}$ which was specific to fact-thrifty chase steps. We show instead that $J' \models \Sigma_{\text{FD}}$, using the assumption that $J \models \Sigma_{\text{FD}}$. Recall the definition of OVL (Definition F.1), and assume by contradiction the existence of a violation of Σ_{FD} in J' . The violation must be between F_n and an existing fact $F = S(\mathbf{c})$. However, because only the elements at positions in O already occur at their position, we must have $\text{OVL}(F_n, F) \subseteq O$. As $\pi_O(F_n)$ was defined using elements of $\text{dom}(\mathcal{E}(D))$, taking $S^r \in \text{OVL}(F_n, F) \subseteq O$, we have $c_r = b_r \in \pi_{S^r}(\mathcal{E}(D))$, so that, by definition of $\mathcal{E}(D)$, we know that $\pi_O(\mathbf{c})$ is a tuple of $\mathcal{E}(D)$. If $\text{OVL}(F_n, F'') \subsetneq O$ then we have a contradiction by applying Lemma F.4 to \mathbf{t} and $\pi_O(\mathbf{c})$ in $\mathcal{E}(D)$. Hence $\text{OVL}(F_n, F'') = O$. So, if D is unsafe, we have a contradiction because F witnesses that \mathbf{t} was not a remaining tuple, so we cannot have used it to define F_n . If D is safe, there is no FD $R^L \rightarrow R^r$ of Σ_{FD} with $R^L \subseteq O$ and $R^r \not\subseteq O$, so F and F_n cannot violate Σ_{FD} , a contradiction again.

We now prove that \mathcal{E} is still a global envelope of J' after performing an envelope-thrifty chase step. The condition on the disjointness of the envelope domains only concerns \mathcal{E} , which is unchanged. Hence, we need only show that, for any $D' \in \text{AFactCl}$, $\mathcal{E}(D')$ is still an envelope. Except the last one, all conditions of the definition of envelopes either concern only the envelope $\mathcal{E}(D')$, which is unchanged, or they are preserved when more facts are created in J' . The last condition needs only to be checked about the new fact F_n created in this chase step.

Except for the elements of F_n at positions in O , all elements of F_n did not occur at the positions where they occur in F_n , by definition of a thrifty chase step. So they cannot be elements of $\text{dom}(\mathcal{E})$ occurring in F_n at the one position where they occur in the one envelope where they occur, because we know that elements from any envelope already occur in J at that position. So we only need to check the condition for the b_r for $S^r \in O$. But because the envelopes of \mathcal{E} are pairwise disjoint and as the b_r are all in $\text{dom}(\mathcal{E}(D))$, we only need to check the condition for $\mathcal{E}(D)$. Now, \mathbf{t} witnesses that $\pi_O(\mathbf{b}) \in \mathcal{E}(D)$. Hence \mathcal{E} is still a global envelope of J' .

Last, to see that the resulting J' is $(n-1)$ -envelope-saturated, it suffices to observe that the new fact F_n witnesses that, for each unsafe class $D \in \text{AFactCl}$, the remaining tuples of $\mathcal{E}(D)$ for J' are those of $\mathcal{E}(D)$ for J minus at most one tuple (namely, some projection of F_n). This concludes the proof.

F.4. Proof of the Envelope-Thrifty Completion Proposition (Proposition VII.10)

Proposition VII.10 (Envelope-thrifty completion). *For any envelope-saturated aligned superinstance J of I_0 that satisfies Σ_{FD} , we can obtain by envelope-thrifty chase steps an aligned superinstance J' of I_0 , such that J' is either envelope-exhausted or satisfies Σ .*

The completion process for envelope-thrifty chase steps is defined in the same way as for fact-thrifty chase steps, except that the elements reused at non-dangerous positions are different. By definition

of thrifty chase steps, the choice of elements reused at those positions cannot make any new UID applicable, or satisfy any UID, because the elements thus reused are required to already occur at the positions where they are used in the new fact. Further, envelope-thrifty chase steps do not introduce UFD violations (in fact, they do not introduce FD violations), as follows from Lemma VII.9. Hence, we can indeed define the completion process for envelope-thrifty chase steps exactly like the completion process for fact-thrifty chase steps, as long as the instance is envelope-saturated. Whenever an envelope-exhausted instance is obtained at any point of the process, we abort and set it to be the final instance.

Assuming that we do not reach any envelope-exhausted instance, the fact that \mathcal{E} is still a global envelope of the result J' of the envelope-thrifty completion process, and that J' satisfies Σ_{FD} in addition to Σ_{UID} , is by Lemma VII.9.

F.5. Proof of the Envelope Blowup Lemma (Lemma VII.11)

Lemma VII.11 (Envelope blowup). *There exists $B \in \mathbb{N}$ depending only on k and Σ_{U} such that, for any aligned superinstance $J = (I, \text{sim})$ of I_0 , and global envelope \mathcal{E} , letting $J' = (I', \text{sim}')$ be the result of the envelope-thrifty completion process, we have $|I'| < B|I|$.*

We first observe that applying a chase round to an aligned superinstance $J = (I, \text{sim})$ of I_0 by any form of thrifty chase steps (Definition V.8) only increases its size by a multiplicative constant. This is because $|\text{dom}(I)| \leq |\sigma| \cdot |I|$, and the number of facts created per element of I in a chase round is at most $|\text{Pos}(\sigma)|$.

Remember that the envelope-completion process starts by constructing an ordered partition $\mathbf{P} = (P_1, \dots, P_n)$ of Σ_{UID} (Definition VI.1). This \mathbf{P} does not depend on the aligned superinstance. Hence, as we satisfy the UIDs of each P_i in turn, if we can show that the instance size only increases by a multiplicative constant for each class, then the blow-up for the entire process is by a multiplicative constant (obtained as the product of the constants for each P_i).

For trivial classes, we apply one chase round by fresh envelope-thrifty chase steps (Corollary VI.4), so the blowup is by a multiplicative constant by our initial observation.

For non-trivial classes, we apply the Fact-Thrifty Completion Proposition (Proposition V.10), modified to use envelope-thrifty rather than fact-thrifty chase steps (but the exact same steps are applied). Remember that this proposition first ensures k -reversibility by applying $k + 1$ envelope-thrifty chase rounds (Proposition D.4) and then makes the result satisfy Σ_{UID} using the Guided Chase Lemma (Lemma D.5). Ensuring k -reversibility only implies a blowup by a multiplicative constant, because it means applying $k + 1$ envelope-thrifty chase rounds. Hence, we focus on the Guided Chase Lemma.

The lemma starts by constructing a balanced pssinstance P using the Balancing Lemma (Lemma IV.9), and a Σ_{U} -compliant piecewise realization PI of P by the Realizations Lemma (Lemma IV.16), and then performs envelope-thrifty chase steps to satisfy Σ_{UID} following PI . We know that, whenever we apply an envelope-thrifty chase step to an element a in the guided chase, a occurs after the chase step at a new position where it did not occur before. Hence, it suffices to show that $|\text{dom}(P)|$ is within a constant factor of $|J|$, because then we know that the final number of facts once the guided chase is over will be $\leq |\text{dom}(P)| \cdot |\text{Pos}(\sigma)|$.

To show this, remember that $\text{dom}(P) = \text{dom}(J) \sqcup \mathcal{H}$, where \mathcal{H} is the helper set. Hence, we only need to show that $|\mathcal{H}|$ is within a multiplicative constant factor of $|J|$. From the proof of the Balancing Lemma, we know that \mathcal{H} is a disjoint union of $\leq |\text{Pos}(\sigma)|$ sets whose size is linear in $|\text{dom}(J)|$ which is itself $\leq |\sigma| \cdot |J|$. Hence, the Guided Chase Lemma only gives rise to a blowup by a constant factor. As we justified, this implies the same about the entire completion process, and concludes the proof.

G. Proofs for Section VIII: Cyclic Queries

In this section, we extend our construction of superinstances that satisfy Σ and are k -sound for ACQ, to superinstances that are k -sound for CQ while still satisfying Σ .

G.1. Proof of the Simple Product Lemma (Lemma VIII.5)

Lemma VIII.5 (Simple product). *Let I be a finite superinstance of I_0 and G a finite $(2k+1)$ -acyclic group generated by $\Lambda(I)$. If I is k -sound for ACQ and k -instance-sound, then $(I, I_0) \otimes G$ is k -sound for CQ.*

Fixing the superinstance I of I_0 that is k -sound for ACQ and k -instance-sound, and the $(2k+1)$ -acyclic group G generated by $\Lambda(I)$, consider $I' := (I, I_0) \otimes G$, which is a superinstance of I_0 (up to our identification of (a, e) to a for $a \in \text{dom}(I_0)$, where e is the neutral element of G). We must show that I' is k -sound for CQ.

We start by proving a simple lemma:

Lemma G.1. *For any CQ q and instance I , if $I \models q$ and some match h of q in I maps two different atoms of q to the same fact F , then there is a strictly smaller q' which entails q and has a match h' in I such that, seeing matches as subinstances of I , $\text{dom}(h') \subseteq \text{dom}(h)$.*

Proof. Fix q , I , h , and let $A = R(\mathbf{x})$ and $A' = R(\mathbf{y})$ be the two atoms of q mapped to the same fact F by h . Necessarily A and A' are atoms for the same relation R of the fact F , and as $h(A) = h(A')$ we know that $h(x_i) = h(y_i)$ for all $R^i \in \text{Pos}(R)$.

Let $\text{dom}(q)$ be the set of variables occurring in q . Consider the application f from $\text{dom}(q)$ to $\text{dom}(q)$ defined by $f(y_i) = x_i$ for all i , and $f(x) = x$ if x does not occur in A' . Observe that this ensures that $h(x) = h(f(x))$ for all $x \in \text{dom}(q)$. Let $q' = f(q)$ be the query obtained by replacing every variable x in q by $f(x)$, and, as $f(A') = f(A)$, removing one of those duplicate atoms so that $|q'| < |q|$. Let $h' = h|_{\text{dom}(q')}$. Clearly the image of h' is a subset of that of h , and to see why this is a match of q' observe that any atom $f(A'')$ of q' is homomorphically mapped by h' to $h(A'')$ because $h'(f(x)) = h(x)$ for all x so $h'(f(A'')) = h(A'')$.

To see why q' entails q , observe that f defines a homomorphism from q to q' , so that, for any match h'' of q' on an instance I' , $h'' \circ f$ is a match of q on I' . \square

Fix now a CQ q such that $|q| \leq k$, and assume that $I' \models q$: let h be a match of q in I . Let us show that $\text{Chase}(I_0, \Sigma_{\text{UID}}) \models q$.

Let pr be the application from I' to I defined by $\text{pr} : (a, g) \mapsto a$ for all $a \in \text{dom}(I)$ and $g \in G$. It is clear that pr is a homomorphism from I' to I that maps $\text{dom}(I_0) \times G$ to $\text{dom}(I_0)$. Hence, if h involves some element of $\text{dom}(I_0) \times G$, then q has a match in I involving an element of I_0 . Hence, as I is k -instance-sound, $\text{Chase}(I_0, \Sigma_{\text{UID}}) \models q$. We accordingly assume that h does not involve an element of $\text{dom}(I_0) \times G$.

If we can show that there is a query q' of ACQ, $|q'| \leq k$, such that q' entails q and $I \models q'$, then, as I is k -sound for ACQ, this suffices to conclude that $\text{Chase}(I_0, \Sigma_{\text{UID}}) \models q'$, hence $\text{Chase}(I_0, \Sigma_{\text{UID}}) \models q$ because q' entails q . So by way of contradiction we assume that q is a query with a match in I' involving no element of $\text{dom}(I_0) \times G$ such that there is no $q' \in \text{ACQ}$, $|q'| \leq k$, where q' entails q and $I \models q'$; and we take this counterexample query q to be of minimal size.

In particular, this means we assume that q is not in ACQ, otherwise we could take $q' = q$, because $I \models q$, as evidenced by $\text{pr} \circ h$. So consider a Berge cycle C of q , of the form $A_1, x_1, A_2, x_2, \dots, A_n, x_n,$

where the A_i are pairwise distinct atoms and the x_i pairwise distinct variables, and for all $1 \leq i \leq n$, variable x_i occurs at position q_i of atom A_i and position p_{i+1} of A_{i+1} , with addition modulo $n := |C|$. We assume without loss of generality that $p_i \neq q_i$ for all i . However, we do not assume that $n \geq 2$: either $n \geq 2$ and C is really a Berge cycle according to our previous definition, or $n = 1$ and variable x_1 occurs in atom A_1 at positions $p_1 \neq q_1$, which corresponds to the case where there are multiple occurrences of the same variable in an atom.

For $1 \leq i \leq n$, we write $F_i = R_i(\mathbf{a}^i)$ the image of A_i by h in I' ; by definition of I' , because h involves no element of $I_0 \times G$ and hence no fact of $I_0 \times G$, there is a fact $F'_i = R_i(\mathbf{b}^i)$ of I and $g_i \in G$ such that $\mathbf{a}^i_j = (b^i_j, g_i \cdot l^i_j)$ for $R^i_j \in \text{Pos}(R_i)$. Now, for all $1 \leq i \leq n$, as $h(x_i) = \mathbf{a}^i_{q_i} = \mathbf{a}^{i+1}_{p_{i+1}}$ for all $1 \leq i \leq n$, we deduce by projecting on the second component that $g_i \cdot l^{F'_i}_{q_i} = g_{i+1} \cdot l^{F'_{i+1}}_{p_{i+1}}$, so that, by collapsing the equations of the cycle together, $l^{F'_1}_{q_1} \cdot (l^{F'_2}_{p_2})^{-1} \cdot \dots \cdot l^{F'_{n-1}}_{q_{n-1}} \cdot (l^{F'_n}_{p_n})^{-1} \cdot l^{F'_n}_{q_n} \cdot (l^{F'_1}_{p_1})^{-1} = e$.

As the girth of G under $\Lambda(I)$ is $\geq 2k + 1$, and this product contains $2n \leq 2k$ elements, we must have either $l^{F'_i}_{q_i} = l^{F'_{i+1}}_{p_{i+1}}$ for some i , or $l^{F'_i}_{p_i} = l^{F'_i}_{q_i}$ for some i . The second case is impossible because we assumed that $p_i \neq q_i$ for all $1 \leq i \leq n$. Hence, necessarily $l^{F'_i}_{q_i} = l^{F'_{i+1}}_{p_{i+1}}$, so in particular $F'_i = F'_{i+1}$. Hence the atoms $A_i \neq A_{i+1}$ of q are mapped by h to the same fact $F'_i = F'_{i+1}$. We conclude by Lemma G.1 that there is a strictly smaller q' which entails q and has a match in I' which is a submatch of h ; so in particular it involves no element of $\text{dom}(I_0) \times G$. Now, by minimality of q , q' cannot be a counterexample query. So there is $q'' \in \text{ACQ}$, $|q''| \leq k$, where q'' entails q' and $I \models q''$. Now, as q'' entails q' and q' entails q , then q'' entails q , so this contradicts the fact that q was a counterexample.

Hence, there is no such counterexample query q , and I' is indeed k -sound for CQ. This concludes the proof.

G.2. Proof of Lemma VIII.8 (Lifting k -bounded simulations to the quotient)

Lemma VIII.8. *Any k -bounded simulation from an instance I to an instance I' defines a k -bounded simulation from I/\simeq_k to I' .*

Fix the instance I and the k -bounded simulation sim to an instance I' , and consider $I'' := I/\simeq_k$. We show that there is a k -bounded simulation sim' from I'' to I' , because $\text{sim} \circ \text{sim}'$ would then be a k -bounded simulation from I'' to I' , the desired claim. We define $\text{sim}'(A)$ for all $A \in I''$ to be a for any member $a \in A$ of the equivalence class A , and show that sim' thus defined is indeed a k -bounded simulation.

We will show the stronger result that $(I'', A) \leq_k (I, a)$ for all $A \in \text{dom}(I'')$ and for any $a \in A$. We do it by proving, by induction on $0 \leq k' \leq k$, that $(I'', A) \leq_{k'} (I, a)$ for all $A \in \text{dom}(I'')$ and $a \in A$. The case $k' = 0$ is trivial. Hence, fix $0 < k' \leq k$, assume that $(I'', A) \leq_{k'-1} (I, a)$ for all $A \in \text{dom}(I'')$ and $a \in A$, and show that this is also true for k' . Choose $A \in \text{dom}(I'')$, $a \in A$, and show that $(I'', A) \leq_{k'} (I, a)$. To do so, consider any fact $F = R(A)$ of I'' such that $A_p = A$ for some $R^p \in \text{Pos}(R)$. Let $F' = R(\mathbf{a}')$ be a fact of I that is a preimage of F by χ_{\simeq_k} , so that $\mathbf{a}'_q \in A_q$ for all $R^q \in \text{Pos}(R)$. We have $\mathbf{a}'_p \in A$ and $a \in A$, so that $\mathbf{a}'_p \simeq_k a$ holds in I . Hence, in particular we have $(I, \mathbf{a}'_p) \leq_{k'} (I, a)$ because $k' \leq k$, so there exists a fact $F'' = R(\mathbf{a}'')$ of I such that $\mathbf{a}''_p = a$ and $(I, \mathbf{a}'_q) \leq_{k'-1} (I, \mathbf{a}''_q)$ for all $R^q \in \text{Pos}(R)$. We show that F'' is a witness fact for F . Indeed, we have $\mathbf{a}''_p = a$. Let us now choose $R^q \in \text{Pos}(R)$ and show that $(I'', A_q) \leq_{k'-1} (I, \mathbf{a}''_q)$. By induction hypothesis, as $\mathbf{a}'_q \in A_q$, we have $(I'', A_q) \leq_{k'-1} (I, \mathbf{a}'_q)$, and as $(I, \mathbf{a}'_q) \leq_{k'-1} (I, \mathbf{a}''_q)$, by transitivity we have indeed $(I'', A_q) \leq_{k'-1} (I, \mathbf{a}''_q)$. Hence, we have shown that $(I'', A) \leq_{k'} (I, a)$.

By induction, we conclude that $(I'', A) \leq_k (I, a)$ for all $A \in \text{dom}(I'')$ and $a \in A$, so that there is indeed a k -bounded simulation from I'' to I , which, as we have explained, implies the desired claim.

G.3. Proof of the Cautiousness Lemma (Lemma VIII.10)

Lemma VIII.10 (Cautiousness). *The superinstance I_f of I_0 constructed by the Acyclic Universal Models Theorem (Theorem VII.1) is cautious for χ_{\simeq_k} .*

Let $J_f = (I_f, \text{sim})$ be the aligned superinstance of I_0 constructed by the Acyclic Universal Models Theorem (Theorem VII.1), and show that it is cautious for χ_{\simeq_k} .

We first observe that the definition of cautiousness (Definition VIII.9) can be generalized to apply to any function, and not just homomorphisms. In this case, writing $F = R(\mathbf{a})$ and $F' = R(\mathbf{a}')$, we define cautiousness as requiring, instead of $h(F) = h(F')$, that $h(a_i) = h(a'_i)$ for all $1 \leq i \leq |R|$.

Now, let χ'_{\simeq_k} be the homomorphism from $\text{Chase}(I_0, \Sigma_{\text{UID}})$ to its quotient by \simeq_k . (We distinguish it from χ_{\simeq_k} , which is the homomorphism from I_f to I_f/\simeq_k .) We first show that our construction ensures the following:

Lemma G.2. *I_f is cautious for $\chi'_{\simeq_k} \circ \text{sim}$.*

In other words, whenever two facts $F = R(\mathbf{a})$ and $F' = R(\mathbf{b})$ overlap in I_f and are not both in I_0 , then, for any position $R^p \in \text{Pos}(R)$, we have $\text{sim}(a_p) \simeq_k \text{sim}(b_p)$ in $\text{Chase}(I_0, \Sigma_{\text{UID}})$.

Proof. In the proof of the Acyclic Universal Models Theorem (Theorem VII.1), I_f is constructed by first constructing an instance I using the Sufficiently Envelope-Saturated Solutions Proposition (Proposition VII.5), and then completing I using the Envelope-Thrifty Completion Proposition (Proposition VII.10).

Thus, we first check that this claim holds for I . Indeed, we check it for each instance constructed in Lemma F.6, and the only overlapping facts in each such instance which are not in I_0 are the $h(F_i)$, which all map to \simeq_k -equivalent sim-images. Hence, as I is the disjoint union of the instances constructed in Lemma F.6, we deduce that the claim holds for I .

Second, in the proof of the Envelope-Thrifty Completion Proposition, we only perform envelope-thrifty chase steps. By their definition, whenever we create a new fact F_n for a fact class D , the only elements of F_n that can be part of an overlap between F_n and an existing fact are envelope elements, appearing at the one position at which they appear in $\mathcal{E}(D)$. Then, by the last condition in the definition of envelopes (Definition VII.2), we deduce that the two overlapping facts achieve the same fact class, which is what we wanted to show. \square

We now want to show that two elements in J_f having \simeq_k -equivalent sim images in $\text{Chase}(I_0, \Sigma_{\text{UID}})$ must themselves be \simeq_k -equivalent in J_f . We do it by showing that, in fact, for any $a \in \text{dom}(J_f)$, not only do we have $(I_f, a) \leq_k (\text{Chase}(I_0, \Sigma_{\text{UID}}), \text{sim}(a))$, but we also have the reverse: $(\text{Chase}(I_0, \Sigma_{\text{UID}}), \text{sim}(a)) \leq_k (I_f, a)$. In other words, intuitively, the facts of the chase must be “mirrored” in I_f .

We define the **ancestry** \mathcal{A}_F of a fact F in $\text{Chase}(I_0, \Sigma_{\text{UID}})$ as I_0 plus the facts of the path in the chase forest that leads to F (if $F \in I_0$ then \mathcal{A}_F is just I_0). The **ancestry** \mathcal{A}_a of $a \in \text{dom}(\text{Chase}(I_0, \Sigma_{\text{UID}}))$ is that of the fact where a was introduced.

We now claim the following:

Lemma G.3. *For any $a \in \text{dom}(I_f)$, there is a homomorphism h_a from $\mathcal{A}_{\text{sim}(a)}$ to I_f such that $h_a(\text{sim}(a)) = a$.*

Proof. We prove that this property holds on I_f , by first showing that it is true of the instance constructed in the Sufficiently Envelope-Saturated Solutions Proposition (Proposition VII.5). This is clearly the case because the instances created by Lemma F.6 are just truncations of the chase where some elements are identified.

Second, we show that the property is maintained by the construction of the Envelope-Thrifty Completion Proposition. We show the stronger claim that it is preserved by any thrifty chase step (Definition V.8). Consider a thrifty chase step where, in a state $J_1 = (I_1, \text{sim}_1)$ of the construction of our aligned superinstance, we apply a UID $\tau : R^p \subseteq S^q$ to a fact $F_a = R(\mathbf{a})$ to create a fact $F_n = S(\mathbf{b})$ and obtain the aligned superinstance $J_2 = (I_2, \text{sim}_2)$. Consider the chase witness $F_w = S(\mathbf{b}')$. By Lemma D.2, b'_q is the exported element between F_w and its parent in $\text{Chase}(I_0, \Sigma_{\text{UID}})$. So we know that for any $i \neq q$, we have $\mathcal{A}_{b'_i} = \mathcal{A}_{b'_q} \sqcup \{F_w\}$.

We need to show that the property holds for the b_i that are fresh (otherwise we already know that the property is satisfied, as adding more facts cannot violate the property in J_2 on an element for which it held in J_1). So, if none of the b_i are fresh, there is nothing to do. Otherwise, choose i such that b_i is fresh. By the definition of thrifty chase steps, we have set $\text{sim}(b_i) := b'_i$. Because $a_p = b_q$ is in $\text{dom}(I_1)$, we know that there is a homomorphism h_{b_q} from $\mathcal{A}_{\text{sim}(b_q)} = \mathcal{A}_{b'_q}$ to I_1 such that we have $h(b'_q) = b_q$. We extend h_{b_q} to the homomorphism h_{b_i} from $\mathcal{A}_{b'_i} = \mathcal{A}_{b'_q} \sqcup \{F_w\}$ to I_2 such that $h_{b_i}(b'_i) = b_i$, by setting $h_{b_i}(F_w) := F_n$ and $h_{b_i}(F) := h(F)$ for any other F of $\mathcal{A}_{b'_i}$; we can do this because, by definition of the chase, F_w shares no element with the other facts of $\mathcal{A}_{b'_i}$ (that is, with $\mathcal{A}_{b'_q}$), except b'_q for which our definition coincides with the existing image. This proves the claim. \square

We claim that this property implies the following:

Corollary G.4. *For any $a \in \text{dom}(I_f)$, there is a homomorphism h_a from $\text{Chase}(I_0, \Sigma_{\text{UID}})$ to I_f such that $h_a(\text{sim}(a)) = a$.*

Proof. Choose $a \in \text{dom}(I_f)$ and let us construct h_a . Let h'_a be the homomorphism from $\mathcal{A}_{\text{sim}(a)}$ to I_f with $h'_a(\text{sim}(a)) = a$ whose existence was proved in Lemma G.3. Now start by setting $h_a := h'_a$, and extend h'_a to be the desired homomorphism, fact by fact, using the property that $I_f \models \Sigma_{\text{UID}}$: for any $b \in \text{dom}(\text{Chase}(I_0, \Sigma_{\text{UID}}))$ not in the domain of h'_a but which was introduced in a fact F whose exported element c is in the current domain of h'_a , let us extend h'_a to the elements of F in the following way: consider the parent fact F' of F and its match by h'_a , let τ be the UID used to create F' from F , and, because $I_f \models \tau$, there must be a suitable fact F'' to extend h'_a to all elements of F by setting $h'_a(F) := F''$; this is consistent with the image of c previously defined in h'_a . Performing this process allows us to define the desired homomorphism h_a . \square

Clearly this result implies:

Corollary G.5. *For any $a \in \text{dom}(I_f)$, we have $(\text{Chase}(I_0, \Sigma_{\text{UID}}), \text{sim}(a)) \leq_k (I_f, a)$.*

Proof. Consider the restriction of h_a to the neighborhood at distance k in the Gaifman graph of $\text{sim}(a)$. \square

We are now ready to show our desired claim:

Lemma G.6. *For any $a, b \in \text{dom}(I_f)$, if $\text{sim}(a) \simeq_k \text{sim}(b)$ in $\text{Chase}(I_0, \Sigma_{\text{UID}})$, then $a \simeq_k b$ in I_f .*

Proof. Fix $a, b \in \text{dom}(I_f)$. We have $(I_f, a) \leq_k (\text{Chase}(I_0, \Sigma_{\text{UID}}), \text{sim}(a))$ because sim is a k -bounded simulation; we have $(\text{Chase}(I_0, \Sigma_{\text{UID}}), \text{sim}(a)) \leq_k (\text{Chase}(I_0, \Sigma_{\text{UID}}), \text{sim}(b))$ because $\text{sim}(a) \simeq_k \text{sim}(b)$; and we have $(\text{Chase}(I_0, \Sigma_{\text{UID}}), \text{sim}(b)) \leq_k (I_f, b)$ by Corollary G.5. By transitivity, we have $(I_f, a) \leq_k (I_f, b)$. The other direction is symmetric, so the desired claim follows. \square

We prove Lemma VIII.8 immediately from Lemma G.2 and Lemma G.6.

G.4. Proof of the Mixed Product Preservation Lemma (Lemma VIII.12)

Lemma VIII.12 (Mixed product preservation). *For any UID or FD τ , if $I \models \tau$ and I is cautious for h , then $(I, I_0) \otimes^h G \models \tau$.*

Write $I_m := (I, I_0) \otimes^h G$.

If τ is a UID, the claim is immediate even without the cautiousness hypothesis. (In fact, the analogous claim could even be proven for the simple product.) Indeed, for any $a \in \text{dom}(I)$ and $R^p \in \text{Pos}(\sigma)$, if $a \in \pi_{R^p}(I)$ then $(a, g) \in \pi_{R^p}(I_m)$ for all $g \in G$; conversely, if $a \notin \pi_{R^p}(I)$ then $(a, g) \notin \pi_{R^p}(I_m)$ for all $g \in G$. Hence, letting $\tau : R^p \subseteq S^q$ be a UID of Σ_{UID} , if there is $(a, g) \in \text{dom}(I_m)$ such that $(a, g) \in \pi_{R^p}(I_m)$ but $(a, g) \notin \pi_{S^q}(I_m)$ then $a \in \pi_{R^p}(I)$ but $a \notin \pi_{S^q}(I)$. Hence any violation of τ in I_m implies the existence of a violation of τ in I , so we conclude because $I \models \tau$.

Assume now that τ is a FD $\phi : R^L \rightarrow R'$. Assume by contradiction that there are two facts $F_1 = R(\mathbf{a})$ and $F_2 = R(\mathbf{b})$ in I_m that violate ϕ , i.e., we have $a_l = b_l$ for all $l \in L$, but $a_r \neq b_r$. Write $a_i = (v_i, f_i)$ and $b_i = (w_i, g_i)$ for all $R^i \in \text{Pos}(R)$. Consider $F'_1 := R(\mathbf{v})$ and $F'_2 := R(\mathbf{w})$ the facts of I that are the images of F_1 and F_2 by the homomorphism from I_m to I that projects on the first component. As $I \models \tau$, F'_1 and F'_2 cannot violate ϕ , so as $v_l = w_l$ for all $l \in L$, we must have $v_r = w_r$. Further, we have $\pi_{R^{l_0}}(F'_1) = \pi_{R^{l_0}}(F'_2)$ for any $l_0 \in L$; hence, as I is cautious for h , either $F'_1, F'_2 \in I_0$ or $h(F'_1) = h(F'_2)$.

In the first case, by definition of the mixed product, there are $f, g \in G$ such that $f_i = f$ and $g_i = g$ for all $R^i \in \text{Pos}(R)$. Thus, taking any $l_0 \in L$, as we have $a_{l_0} = b_{l_0}$, we have $f_{l_0} = g_{l_0}$, so $f = g$, which implies that $f_r = g_r$. Hence, as $v_r = w_r$, we have $(v_r, f_r) = (w_r, g_r)$, contradicting the fact that $a_r \neq b_r$.

In the second case, as h is the identity on I_0 and maps $I \setminus I_0$ to $I' \setminus I_0$, $h(F'_1) = h(F'_2)$ implies that either F'_1 and F'_2 are both facts of I_0 or they are both facts of $I \setminus I_0$; but we have already excluded the former possibility in the first case, so we assume the latter. Let F be $h(F'_1)$. By definition of the mixed product, there are $f, g \in G$ such that $f_i = f \cdot 1_i^{h(F)}$ and $g_i = g \cdot 1_i^{h(F)}$ for all $R^i \in \text{Pos}(R)$. Picking $l_0 \in L$, from $a_{l_0} = b_{l_0}$, we deduce that $f \cdot 1_{l_0}^{h(F)} = g \cdot 1_{l_0}^{h(F)}$, which simplifies to $f = g$. Hence, $f_r = g_r$ and we conclude like in the first case.

G.5. Proof of the Mixed Product Homomorphism Lemma (Lemma VIII.13)

Lemma VIII.13 (Mixed product homomorphism). *There is a homomorphism from $(I, I_0) \otimes^h G$ to $(I', I_0) \otimes G$ which is the identity on $I_0 \times G$.*

We use the homomorphism $h : I \rightarrow I_1$ to define the homomorphism h' from $I_m := (I, I_0) \otimes^h G$ to $I_p := (I, I_0) \otimes G$ by $h'((a, g)) := (h(a), g)$ for every $(a, g) \in \text{dom}(I) \times G$.

Consider a fact $F = R(\mathbf{a})$ of I_m , with $a_i = (v_i, g_i)$ for all $R^i \in \text{Pos}(R)$. Consider its image $F' = R(\mathbf{v})$ by the homomorphism from I_m to I obtained by projecting to the first component, and the image $h(F')$ of F' by the homomorphism h . As $h|_{I_0}$ is the identity and $h|_{(I \setminus I_0)}$ maps to $I_1 \setminus I_0$, $h(F')$ is a fact of I_0 iff F' is. Now by definition of the simple product it is clear that I_p contains the fact $h'(F)$ (it was created in I_p from $h(F')$ for the same choice of $g \in G$).

The fact that h is the identity on I_0 also ensures that h' is the identity on $I_0 \times G$.